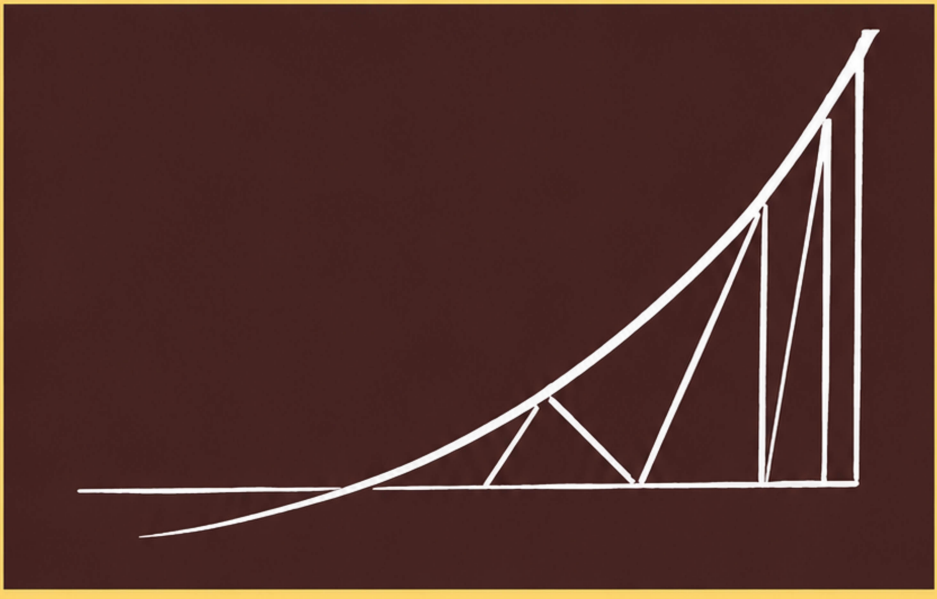
V.A. ILYIN and E.G. POZNYAK

Fundamentals of Mathematical Analysis

PART 1

Mir Publishers Moscow



V. A. Ilyin and E. G. Poznyak

Fundamentals of Mathematical Analysis

PART 1

This course of mathematical analysis, which can be recommended to first-year University students, includes the theory of real numbers, the theory of limits, differential and integral calculus of functions in one variable, the theory of number series, differential calculus of functions in several variables, the theory of implicit functions, and geometrical applications of the analysis.

A peculiarity of this textbook is the presence of an introductory prepaedeutic chapter, in which the class of the basic notions and ideas constituting the course of mathematical analysis is elucidated in the process of considering a simple problem of a particle moving along a line. Another peculiarity of the book is the exposition on the basis of the approach of mathematical analysis, of algorithms for computing the values of elementary functions and definite integrals and algorithms for calculating extrema based on the gradient methods. The exposition of the theory of the envelope and the discriminant curve is more complete here than in other study aids. The Appendix is devoted to the axiomatic theory of real numbers and to the proof of the theorem of isomorphism and its various realizations.

The textbook was awarded the 1980 State Prize of the USSR.

V.A. ILYIN and E.G. POZNYAK

Fundamentals of Mathematical Analysis

Translated from the Russian
by Irene Aleksanova

PART 1

MIR PUBLISHERS MOSCOW

First published 1982

На английском языке

© Изд-во «Наука», Главная редакция физико-математической литературы, 1982

© English translation, Mir Publishers, 1982

В. А. ИЛЬИН, Э. Г. ПОЗНЯК

ОСНОВЫ МАТЕМАТИЧЕСКОГО АНАЛИЗА

Часть I

Издательство «Наука» Москва

ИЗДАТЕЛЬСТВО
МИР

PREFACE TO THE RUSSIAN EDITION

The book is based on the lectures delivered by the authors at the Physics Department of Moscow University.

In the course of systematic exposition of the material, the authors gave prominence to the most important concepts and theorems, the most significant of the latter being called fundamental theorems. The authors tried to give the formulations of new concepts and theorems not long before their direct application.

The sequence of presentation of the material corresponds to that accepted at the Physics Department of Moscow University. The chapter "Preliminary Information on the Main Concepts of Mathematical Analysis", in particular, comes before the exposition of the systematic course because this chapter deals with some very important physical problems and discusses mathematical means for their solution. Thus it becomes clear from the very beginning what problems and concepts constitute the subject matter of mathematical analysis. Our lecturing experience shows that such a preliminary elucidation of the range of problems treated in the course of the analysis makes it appreciably easier for the students to understand the basic mathematical concepts.

The authors paid due attention to the ever increasing role of computations and methods of approximation; that is why they tried to use algorithmic form in proving theorems and carrying out calculations wherever it was possible. In Chap. 12, in particular, the main emphasis is laid on the algorithmic aspect of the methods of approximation and only then is the method substantiated.

In addition to the main part of the material, the authors consider it possible to include some sections given in small type.

In writing the book, the authors borrowed some methodical techniques from the course of lectures by N.V. Efimov as well as from the well-known books by E. Goursat, De La Vallée Poussin and F. Franklin.

The authors express their deep gratitude to A.N. Tikhonov and A.G. Sveshnikov for their valuable ideas and instructions and also for their enormous help at all stages of writing this book.

Thanks are also due to I.A. Shishmarev whose editing of the book was very beneficial.

The authors also want to express their gratitude to N.V. Efimov and L.D. Kudryavtsev for a large number of methodical instructions, B.M. Budak and S.V. Fomin for reviewing separate chapters and making their remarks, and B. Khimchenko, P. Zaikin and A. Zolotarev for their help in preparing the manuscript for print.

The authors consider it necessary to point out that to a large extent their pedagogical views were influenced by their talks with I.M. Vinogradov and A.A. Samarsky, to whom they are also very thankful.

V. Ilyin, E. Poznyak

PREFACE TO THE ENGLISH EDITION

The present book is a translation into English of the 4th Russian edition of *Fundamentals of Mathematical Analysis*, Part 1, published by Nauka in 1982.

The previous edition has been revised and expanded: some new material has been introduced and a number of changes made, which were called forth by the desire of the authors to improve the exposition and to reflect the growing role of computational techniques and algorithms.

The authors consider it their pleasant duty to express their gratitude to Irene Aleksanova whose thorough work on the translation of the book helped to eliminate a number of inaccuracies which occured in the Russian edition.

The authors.

CONTENTS

Preface to the Russian Edition 5
Preface to the English Edition 6

Chapter 1. Preliminary Information on the Main Concepts of Mathematical Analysis 17

1.1. Mathematical Concepts Needed for the Description of Motion 17
1.2. Instantaneous Velocity and the New Mathematical Concepts Related to It 19
1.3. The Problem of Reconstructing the Law of Motion from the Velocity and the Related Mathematical Problems 27
1.4. Some Questions Arising in Solving the Problems on Computation of a Path 29
1.5. Concluding Remarks 33

Chapter 2. Theory of Real Numbers 35

2.1. Real Numbers 35
2.1.1. The properties of rational numbers 35
2.1.2. Measuring the segments on the number axis 37
2.1.3. Real numbers and the rule of their comparison 40
2.1.4. Approximation of real numbers by rational numbers 42
2.1.5. Sets of real numbers bounded above and below 43
2.2. Arithmetic Operations on Real Numbers. Basic Properties of Real Numbers 47
2.2.1. Definition of the sum of real numbers 47
2.2.2. Definition of the product of real numbers 49
2.2.3. Properties of real numbers 49
2.2.4. Some frequently employed relations 52
2.3. Some Concrete Sets of Real Numbers 52
Supplement 1 to Chapter 2. Decimal-to-Binary Conversion and Binary-to-Decimal Conversion of Numbers 53
1. Decimal-to-binary conversion of numbers 53
2. Binary-to-decimal conversion of numbers 55
Supplement 2 to Chapter 2. On the Errors in Rounding Off the Numbers in the Number Systems with Even and Odd Bases 55

Chapter 3. Limit of a Sequence 57

3.1. Number Sequences 57
3.1.1. Number sequences and operations on them 57
3.1.2. Bounded and unbounded sequences 58
3.1.3. Infinitely large and infinitesimal sequences 59
3.1.4. Basic properties of infinitesimal sequences 60
3.2. Convergent Sequences and Their Basic Properties 63
3.2.1. The concept of a convergent sequence 63
3.2.2. Basic properties of convergent sequences 64

3.2.3. Limiting process in inequalities 67
3.3. Monotonic Sequences 68
3.3.1. Definition of monotonic sequences 68
3.3.2. Test for convergence of a monotonic sequence 69
3.3.3. Some examples of convergent monotonic sequences 70
3.3.4. Number e 73
3.4. Some Properties of Arbitrary Sequences and Number Sets 74
3.4.1. Subsequences of number sequences 74
3.4.2. Limit points of a sequence 75
3.4.3. Existence of a limit point of a bounded sequence 77
3.4.4. Separation of a convergent subsequence 80
3.4.5. A necessary and sufficient condition of convergence of a sequence 82
3.4.6. Some properties of arbitrary number sets 84
Supplement 1 to Chapter 3. Stolz' Theorem 86
Supplement 2 to Chapter 3. On the Rate of Convergence of a Sequence Approaching $\sqrt{a}$ 90

Chapter 4. The Concept of a Function. The Limiting Value of a Function. Continuity 93

4.1. The Concept of a Function 93
4.1.1. A variable quantity and a function 93
4.1.2. Methods of specification of a function 95
4.2. Tne Concept of the Limiting Value of a Function 96
4.2.1. Definition of the limiting value of a function 96
4.2.2. Arithmetic operations on functions possessing limiting values 99
4.2.3. Comparing infinitesimal and infinitely large functions 100
4.3. The Concept of Continuity of a Function 103
4.3.1. Definition of continuity of a function 103
4.3.2. Arithmetic operations on continuous functions 105
4.3.3. Composite function and its continuity 105
4.4. Some Properties of Monotonic Functions 106
4.4.1. Definition and examples of monotonic functions 106
4.4.2. The concept of an inverse function. Monotonic function having an inverse 106
4.5. Basic Elementary Functions 110
4.5.1. Rational powers of positive numbers 110
4.5.2. Exponential function 112
4.5.3. Logarithmic function 116
4.5.4. Hyperbolic functions 117
4.5.5. Power function with any real exponent α 118
4.5.6. Trigonometric functions 121
4.5.7. Inverse trigonometric functions 124
4.6. Limits of Some Functions 127
4.6.1. Preliminary remarks 127
4.6.2. The limit of the function $\frac{\sin x}{x}$ at point $x = 0$ (the first remarkable limit) 127
4.6.3. The limit of the function $\left(1+\frac{1}{x}\right)^x$ for $x \to \infty$ (the second remarkable limit) 128
4.7. Continuity and Limiting Values of Some Composite Functions 132

4.7.1. Continuity and limiting values of some composite functions 132
4.7.2. The concept of an elementary function. The class of elementary functions 136
4.8. Classification of Points of Discontinuity of a Function 137
4.8.1. Points of discontinuity of a function and their classification 137
4.8.2. Piecewise-continuous functions 140
Supplement to Chapter 4. Proof of the Assertion Made in 4.5.6 140
1. The proof of the uniqueness 140
2. The proof of the existence 143

Chapter 5. Fundamentals of Differential Calculus 150

5.1. A derivative. Its Physical and Geometrical Interpretations 150
5.1.1. Increment of an argument and of a function. Difference form of the condition of continuity 150
5.1.2. Definition of a derivative 151
5.1.3. Physical interpretation of a derivative 152
5.1.4. Geometrical interpretation of a derivative 152
5.1.5. Right-hand and left-hand derivatives 154
5.1.6. The concept of a derivative of a vector function 154
5.2. The Concept of a Differentiable Function 156
5.2.1. Differentiation of a function at a given point 156
5.2.2. The connection between the concepts of differentiability and continuity of a function 157
5.2.3. The concept of a differential of a function 157
5.3. Rules for Differentiating a Sum, a Difference, a Product and a Quotient 159
5.4. Calculating the Derivatives of a Power Function, Trigonometric Functions and a Logarithmic Function 162
5.4.1. The derivative of a power function with an integral-valued exponent 162
5.4.2. The derivative of the function $y = \sin x$ 162
5.4.3. The derivative of the function $y = \cos x$ 163
5.4.4. The derivatives of the functions $y = \tan x$ and $y = \cot x$ 164
5.4.5. The derivatives of the function $y = \log_a x$ $(0 < a \neq 1)$ 164
5.5. Theorem on the Derivative of an Inverse Function 165
5.6. Computing the Derivatives of an Exponential Function and of Inverse Trigonometric Functions 166
5.6.1. The derivative of the exponential function $y = a^x$ $(0 < a \neq 1)$ 166
5.6.2. The derivatives of inverse trigonometric functions 167
5.7. The Rule of Differentiating a Composite Function 169
5.8. The Logarithmic Derivative. The Derivative of a Power Function with Any Real Exponent. The Table of the Derivatives of Basic Elementary Functions 171
5.8.1. The concept of the logarithmic derivative of a function 171
5.8.2. The derivative of a power function with any real exponent 171
5.8.3. The table of derivatives of basic elementary functions 172
5.9. Invariance of the Form of the First Differential. Some Applications of a Differential 173
5.9.1. Invariance of the form of the first differential 173
5.9.2. Formulas and rules for computing differentials 175
5.9.3. Application of differentials for establishing approximation formulas 175

5.10. Derivatives and Differentials of Higher Orders 177
5.10.1. The concept of a derivative of the nth order 177
5.10.2. Derivatives of the nth order of some functions 177
5.10.3. Leibniz' formula for the nth order derivative of the product of two functions 179
5.10.4. Higher-order differentials 180
5.11. Differentiation of a Function Represented Parametrically 182

Chapter 6. Indefinite Integral 184

6.1. The Concepts of an Antiderivative of a Function and of the Indefinite Integral 184
6.1.1. The concept of an antiderivative of a function 184
6.1.2. Indefinite integral 185
6.1.3. Basic properties of the indefinite integral 186
6.1.4. Table of the basic indefinite integrals 187
6.2. Main Methods of Integration 190
6.2.1. Integration by a change of variable (substitution) 190
6.2.2. Integration by parts 193

Chapter 7. Complex Numbers. Algebra of Polynomials Integration in Terms of Elementary Functions 197

7.1. Brief Information on Complex Numbers 197
7.2. Algebraic Polynomials 201
7.3. Multiple Roots of Polynomials. Test for Distinguishing the Multiplicity of a Root 203
7.4. The Principle of Separation of Multiple Roots. Euclidean Algorithm 205
7.4.1. The principle of separation of multiple roots 205
7.4.2. Finding the greatest common divisor of two polynomials (Euclidean algorithm) 206
7.5. Decomposition of a Proper Rational Fraction with Complex Coefficients into a Sum of Partial Fractions 209
7.6. Factorization of an Algebraic Polynomial with Real Coefficients into a Product of Irreducible Real Factors 211
7.7. Decomposition of a Proper Rational Fraction with Real Coefficients into a Sum of Partial Fractions with Real Coefficients 213
7.8. The Problem of Integrating Rational Fractions 218
7.9. Ostrogradsky's Method 221
7.10. Integrating Some Irrational and Transcendental Expressions 224
7.10.1. Integrating certain trigonometric expressions 224
7.10.2. Integrating linear-fractional irrational functions 227
7.10.3. Integtating binomial differentials 228
7.10.4. Integrating quadratic irrational functions by means of Euler's substitution 230
7.10.5. Integrating quadratic irrational functions by other means 232
7.11. Elliptic Integrals 238

Chapter 8. Fundamental Theorems on Continuous and Differentiable Functions 241

8.1. A New Definition of the Limiting Value of a Function 241
8.1.1. A new definition of the limiting value of a function. Its equivalence to the definition given earlier 241

8.1.2. The necessary and sufficient condition of the existence of a limiting value (Cauchy's test) 244
8.2. A Local Boundedness of a Function Having a Limiting Value 246
8.3. Theorem on the Stability of Sign of a Continuous Function 248
8.4. The Passage by a Continuous Function Through Any Intermediate Value 249
8.4.1. The passage by a continuous function through zero at a change of signs 249
8.4.2. The passage by a continuous function through any intermediate value 250
8.5. The Boundedness of a Function Continuous on a Closed Interval 250
8.6. The Least Upper Bound and the Greatest Lower Bound of a Function and Their Attainment by a Function Continuous on a Closed Interval 251
8.6.1. The concepts of the least upper bound (the supremum) and the greatest lower bound (the infimum) of a function on a given set 251
8.6.2. Attaining by a function, continuous on a closed interval, its least upper and greatest lower bounds 252
8.7. A Function Increasing (Decreasing) at a Point. A Local Maximum (Minimum) 254
8.7.1. An increase (decrease) of a function at a point 254
8.7.2. A local maximum and a local minimum of a function 255
8.8. Theorem on a Zero of a Derivative 256
8.9. The Formula of Finite Increments (Lagrange's Formula) 257
8.10. Some Corollaries of Lagrange's Formula 258
8.10.1. The constant value of the function possessing a zero derivative on an interval 258
8.10.2. Conditions for monotonicity of a function on an interval 259
8.10.3. A derivative possessing no points of discontinuity of the 1st kind or removable discontinuity 261
8.10.4. Derivation of certain inequalities 262
8.11. A Generalized Formula of Finite Increments (Cauchy's Formula) 263
8.12. Evaluation of Indeterminate Forms (L'Hospital's Rule) 264
8.12.1. Evaluating indeterminate expressions of the form $\frac{0}{0}$ 264
8.12.2. Evaluating indeterminate expressions of the form $\frac{\infty}{\infty}$ 266
8.12.3. Evaluating indeterminate expressions of other forms 268
8.13. Taylor's Formula 269
8.14. Various Forms of the Remainder, Maclaurin's Formula 272
8.14.1. The remainder in the form of Lagrange, Cauchy and Peano 272
8.14.2. Another form of Taylor's formula 275
8.14.3. Maclaurin's formula 275
8.15. Evaluation of the Remainder. Expansion of Some Elementary Functions 276
8.15.1. Evaluating the remainder for an arbitrary function 276
8.15.2. Expanding some elementary functions by Maclaurin's formula 277
8.16. Examples of Applications of Maclaurin's Formula 280
8.16.1. Algorithm for calculating the number *e* 280

8.16.2. Realization of algorithm for calculating the number e by means of a computer 280
8.16.3. Application of Maclaurin's formula for asymptotic estimations of elementary functions and for computing limits 282
Supplement to Chapter 8. Calculation of Elementary Functions 285
1. Calculation of the logarithmic function and the inverse trigonometric functions 285
2. Computing the trigonometric functions, the exponential function and the hyperbolic functions 288

Chapter 9. Geometric Investigation of the Graph of a Function. Finding the Maximum and Minimum Values of a Function 295

9.1. Intervals of Monotonicity of a Function. Finding the Points of Extremum 295
9.1.1. Determining the intervals of monotonicity of a function 295
9.1.2. Finding the points of a possible extremum 295
9.1.3. The first sufficient condition of extremum 296
9.1.4. The second sufficient condition of extremum 298
9.1.5. Extremum of a function nondifferentiable at a given point. A general scheme of searching for extrema 301
9.2. The Direction of the Convexity of the Graph of a Function 304
9.3. Point of Inflection of the Graph of a Function 306
9.3.1. Definition of the point of inflection. A necessary condition for inflection 306
9.3.2. The first sufficient condition of inflection 309
9.3.3. The second sufficient condition of inflection 309
9.3.4. Some generalizations of the first sufficient condition of inflection 310
9.4. The Third Sufficient Condition of an Extremum and an Inflection 311
9.5. Asymptotes of the Graph of a Function 317
9.6. The Scheme for Investigating the Graph of a Function 316
9.7. Finding the Maximum and the Minimum Value of a Function. An End-Point Extremum 319
9.7.1. Finding the maximum and minimum value of a function 319
9.7.2. End-point extremum 322

Chapter 10. Definite Integral 324

10.1. Integral Sums. Integrability 324
10.2. Upper and Lower Sums 327
10.2.1. The concept of an upper and a lower sum 327
10.2.2. The properties of the upper and lower sums 328
10.3. A Necessary and Sufficient Condition of Integrability 332
10.4. Some Classes of Integrable Functions 334
10.4.1. The property of uniform continuity of a function 334
10.4.2. Heine-Borel lemma. Another proof of the theorem on uniform continuity 337
10.4.3. Integrability of continuous functions 338
10.4.4. Integrability of some discontinuous functions 339
10.4.5. Integrability of monotonic bounded functions 341
10.5. The Main Properties of the Definite Integral 341
10.6. Evaluations of Integrals. Mean-Value Formulas 344
10.6.1. Evaluations of integrals 344
10.6.2. The first mean-value formula 347

10.6.3. The first mean-value formula in a generalized form 348
10.6.4. The second mean-value formula 349
10.7. The Existence of an Antiderivative of a Continuous Function. Basic Rules of Integration 349
10.7.1. The existence of an antiderivative of a continuous function 349
10.7.2. The fundamental formula of integral calculus 351
10.7.3. A change of variable under the definite integral sign 353
10.7.4. The formula of integration by parts 354
10.7.5. The remainder in Taylor's formula in integral form 355
Supplement 1 to Chapter 10. Some Important Inequalities for Sums and Integrals 357
1.1. Derivation of a preliminary inequality 357
1.2. Hölder's inequality for sums 358
1.3. Minkowski's inequality for sums 359
1.4. Integrability of an arbitrary positive degree of the modulus of the function being integrated 359
1.5. Hölder's inequality for integrals 360
1.6. Minkowski's inequality for integrals 362
Supplement 2 to Chapter 10. Proof of the Assertion from 10.6.4 362

Chapter 11. Applications of the Definite Integral in Geometry and Physics 365

11.1. The Length of the Arc of a Curve 365
11.1.1. The concept of a plane curve 365
11.1.2. Parametric representation of a curve 366
11.1.3. The notion of a space curve 368
11.1.4. The concept of an arc length of a curve 369
11.1.5. Sufficient conditions for rectifiability of a curve. Formulas for computing the arc length of a curve 373
11.1.6. Differential of an arc 378
11.1.7. Some examples of arc length computation 379
11.2. Area of a Plane Figure 380
11.2.1. The concept of squareability of a plane figure. Area of a squareable plane figure 380
11.2.2. Area of a curvilinear trapezoid 382
11.2.3. Area of a curvilinear sector 383
11.2.4. Examples of computing areas 385
11.3. Volumes of Bodies and Areas of Surfaces 386
11.3.1. The concept of cubability and volume 386
11.3.2. Cubability of some classes of bodies 387
11.3.3. Examples of computing volumes 389
11.3.4. Area of a surface of revolution 389
11.4. Some Physical Applications of the Definite Integral 392
11.4.1. The mass and the centre of gravity of a nonhomogeneous bar 392
11.4.2. The work performed by a variable force 393
Supplement to Chapter 11. An Example of a Nonsquareable Figure 394

Chapter 12. Approximate Methods of Calculating Roots of Equations and Definite Integrals 399

12.1. Approximate Methods of Calculating Roots of Equations 399
12.1.1. The "fork" method 399
12.1.2. The tangent method 400

12.1.3. The chord method 401
12.1.4. The method of iterations (successive approximations) 402
12.1.5. Substantiation of the tangent method 405
12.1.6. Substantiation of the chord method 409
12.2. Approximate Methods of Calculating Definite Integrals 411
12.2.1. Introductory remarks 411
12.2.2. The method of rectangles 414
12.2.3. The method of trapezoids 417
12.2.4. The method of parabolas 419
12.2.5. Concluding remarks 422

Chapter 13. Theory of Number Series 424

13.1. The Concept of a Number Series 424
13.1.1. A series and its partial sums. Convergent and divergent series 424
13.1.2. Cauchy's test for convergence of a series 427
13.1.3. Two properties connected with the convergence of a series 429
13.2. Series with Positive Terms 429
13.2.1. A necessary and sufficient condition for convergence of a series with positive terms 429
13.2.2. Comparison tests 430
13.2.3. D'Alembert's and Cauchy's tests 433
13.2.4. Cauchy-Maclaurin's integral test 437
13.2.5. Raabe's test 440
13.2.6. Absence of a universal series of comparison 442
13.3. Absolutely and Conditionally Convergent Series 443
13.3.1. The concept of an absolutely and conditionally convergent series 443
13.3.2. On rearrangement of terms of conditionally convergent series 445
13.3.3. On rearrangement of an absolutely convergent series 448
13.4. Arithmetic Operations on Convergent Series 450
13.5. Tests of Convergence of Arbitrary Series 452
13.5.1. Leibniz's test 452
13.5.2. The test of Dirichlet-Abel 454
13.6. Infinite Products 457
13.6.1. Basic notions 457
13.6.2. Connection between the convergence of continued products and series 459
Supplement 1 to Chapter 13. Auxiliary Theorem to 13.2.3 463
Supplement 2 to Chapter 13. Expansion of the Function $\sin x$ into a Continued Product 464
Supplement 3 to Chapter 13. Generalized Methods of Summation of Divergent Series 468

Chapter 14. Functions of Several Variables 472

14.1. The Concept of a Function of Several Variables 472
14.1.1. On functional relationships between several variable quantities 472
14.1.2. The concepts of the Euclidean plane and the Euclidean space 472
14.1.3. The concept of a function of two or three variables 473
14.1.4. The concepts of the m-dimensional coordinate space and m-dimensional Euclidean space 475
14.1.5. The sets of points of the m-dimensional Euclidean space E^m 476

14.1.6. The concept of a function of m variables 478
14.2. The Limiting Value of a Function of Several Variables 479
14.2.1. Convergent sequences of points in m-dimensional Euclidean space E^m. Cauchy's test for convergence of a sequence 479
14.2.2. Some properties of bounded sequences of points in an m-dimensional Euclidean space 481
14.2.3. The concept of the limiting value of a function of several variables 482
14.2.4. Infinitesimal functions 483
14.2.5. A necessary and sufficient condition of the existence of the limiting value of a function (Cauchy's test) 483
14.2.6. Repeated limiting values 484
14.3. Continuous Functions of Several Variables 485
14.3.1. The definition of the continuity of a function of several variables 485
14.3.2. Main properties of discontinuous functions of several variables 488
14.4. Derivatives and Differentials of a Function of Several Variables 491
14.4.1. Partial derivatives of a function of several variables 491
14.4.2. The concept of differentiability of a function of several variables 493
14.4.3. The concept of the differential of a function of several variables 499
14.4.4. Differentiation of a composite function 499
14.4.5. Invariance of the form of the first differential 503
14.4.6. Directional derivative. Gradient 504
14.5. Partial Derivatives and Differentials of Higher Orders 507
14.5.1. Partial derivatives of higher orders 507
14.5.2. Differentials of higher orders 511
14.5.3. Taylor's formula for the function of m variables with the remainder in Lagrange's form 517
14.5.4. Taylor's formula with the remainder in Peano's form 520
14.6. A Local Extremum of a Function of m Variables 524
14.6.1. The concept of an extremum of a function of m variables. Necessary conditions of an extremum 524
14.6.2. Sufficient conditions of a local extremum of a function of m variables 526
14.6.3. The case of a function of two variables 533
14.6.4. Examples of investigating functions for an extremum 534
14.7. The Gradient Method of Seeking an Extremum of a Strongly Convex Function 535
14.7.1. Convex sets and convex functions 536
14.7.2. The existence of a minimum of a strongly convex function and the uniqueness of the minimum of a strictly convex function 542
14.7.3. Seeking a minimum of a strongly convex function 547
Supplement to Chapter 14. On the Choice of the Optimum Partitioning of a Closed Interval for Approximate Calculation of an Integral 555

Chapter 15. The Theory of Implicit Functions and Its Applications 559

15.1. The Concept of the Implicit Function 559
15.2. Theorem on the Existence and Differentiability of an Implicit Function and Some of Its Applications 560

15.2.1. Theorem on the existence and differentiability of an implicit function 560
15.2.2. Calculating partial derivatives of a function specified implicitly 566
15.2.3. Singularities of a surface and of a plane curve 568
15.2.4. Conditions ensuring the existence of an inverse of the function $y = f(x)$ 569
15.3. Implicit Functions Specified by a System of Functional Equations 570
15.3.1. Theorem on the solvability of a system of functional equations 570
15.3.2. Calculating partial derivatives of functions specified implicitly by a system of functional equations 576
15.3.3. One-to-one mapping of two sets of an m-dimensional space 576
15.4. Functional Dependence 577
15.4.1. The concept of a functional dependence. A sufficient condition of independence 577
15.4.2. Functional matrices and their applications 579
15.5. Conditional Extremum 583
15.5.1. The concept of the conditional extremum 583
15.5.2. Lagrange's method of multipliers 586
15.5.3. Sufficient conditions 587
15.5.4. Example 589
Supplement to Chapter 15. Change of Variables 591

Chapter 16. Some Geometric Applications of Differential Calculus 595

16.1. The Envelope and the Discriminant Curve of a One-Parameter Family of Plane Curves 595
16.1.1. Preliminaries 595
16.1.2. One-parameter families of plane curves. Characteristic points of the curves of a family 597
16.1.3. The envelope and the discriminant curve of a one-parameter family of plane curves 599
16.1.4. The envelope and the discriminant surface of a one-parameter family of surfaces 603
16.2. Contact of Plane Curves 604
16.2.1. The notion of the order of contact of plane curves 604
16.2.2. The order of contact of curves which are graphs of functions 605
16.2.3. Sufficient conditions of contact of order n 608
16.2.4. An osculating circle 609
16.3. The Curvature of a Plane Curve 610
16.3.1. The notion of the curvature of a plane curve 610
16.3.2. Formula for calculating curvature 612
16.4. The Evolute and the Involute 615
16.4.1. A normal to a plane curve 615
16.4.2. The evolute and the involute of a plane curve 617

Appendix. Further Development of the Theory of Real Numbers 621

1. Completeness of a set of real numbers 621
2. Axiomatic introduction of the set of real numbers 624
3. Concluding remarks 629

CHAPTER 1

PRELIMINARY INFORMATION ON THE MAIN CONCEPTS OF MATHEMATICAL ANALYSIS

1.1. MATHEMATICAL CONCEPTS NEEDED FOR THE DESCRIPTION OF MOTION

1.1.1. Mathematics is a science whose aim is to study quantitative relationships and space forms of the world around us.

Elementary mathematics is limited to the quantitative relationships and space forms since it deals essentially with *constant quantities* and simple geometric figures (triangles, circles, etc.). The concepts and methods employed in elementary mathematics are insufficient to describe mechanical motion and other time processes. Let us see now what new mathematical notions are needed for that purpose.*

1.1.2. Every process is associated with some *variable quantity*, i.e. a quantity which takes on different values under the conditions of the given process. Moreover, every process is characterized by at least two variable quantities *whose variations are interrelated.*

Let us consider, for example, mechanical motion of a particle in a straight line. This motion is the process of variation of the position of the particle in a straight line with time. The indicated process is associated with two variables, time and the path traversed by the particle from the reference point. To describe the motion being considered, it is necessary to know the distance from the particle to the reference point at each instant of time, that is the *dependence of the path traversed by the particle on time.* In mechanics, such a dependence is called the *law of motion.* In other words, the law of motion is the rule according to which each value of time x is associated with a definite value of the path y traversed by the particle during the time x.

Such a relationship between the variables x and y which puts each value of the variable x into correspondence with a definite value of the variable y is encountered not only in the description of mechanical motion of a particle but of other physical processes as well. Abstracting from the concrete physical meaning of the variables x and y, we arrive at one of the most important mathematical concepts, that of a *function***.

---

* We do not aim at strict formulations but try to outline the class of problems we shall deal with later on.

** The introduction into mathematics of the concept of a function is associated with the name of the eminent English scientist Isaac Newton (1642-1727).

If there is a rule in accordance with which each value of the variable x is associated with a definite value of the variable y, then the variable y is said to be a function of the variable x.

In that case, the variable x is called the *argument* (*independent variable*) of the function under consideration, and the value of y corresponding to the given value of x is called a *particular value* of the function at point x.

A function can be symbolized as follows:

$$y = y(x) \quad \text{or} \quad y = f(x).$$

In this notation the letter f called a characteristic of a function symbolizes the rule formulated above. If different functions are being considered, then different letters are used to denote their characteristics. Note that x and y are not only letters used to designate an argument and a function. For instance, the notation $S = h(t)$ means that the variable S is the function of the argument t, the characteristic of a function being denoted by the letter h.

Both a variable quantity and a function usually have *numerical values*. Therefore, to acquire a deeper knowledge of these concepts, you have to master the theory of real numbers*.

Let us take some examples of functions.

(1) It is known that the path S traversed, during the time t, by an initially stationary particle falling under the force of gravity can be determined by the equation

$$S = \frac{gt^2}{2}.$$

This formula represents the rule which associates each value of the variable t with the value of the variable S, that is it defines S as a function of the argument t.

(2) In accordance with Coulomb's law two unlike unit charges, positioned at a distance r from each other, are attracted with the force

$$F = \frac{c}{r^2},$$

where c is a certain constant. This is also a formalization of the rule associating every value of the variable r with a value of the variable F, i.e. defining F as a function of the argument r.

---

* It should be noted that the concept of a function and that of a number are the so-called *basic concepts*. Each of them can be explained, but every venture to define a basic concept reduces to the replacement of the concept being defined by its equivalent. The reader is sure to be acquainted with the basic concepts from elementary mathematics. They are, for instance, a straight line, a plane.

In the two examples cited above, the rule of the correspondence of the argument and the function was given *by means of a formula*. Such a method of representation of a function is called *analytical*.

There are other methods of representing a function. Here are some of them. Very often, in the practice of physical measurements, a function is given in a *tabular form*, that is the values of the argument and the corresponding values of the function are tabulated. Not infrequently, the relation between an argument and a function is shown by means of a graph which can be recorded by an oscillograph, for instance. This method of representing a function is called *graphical*.*

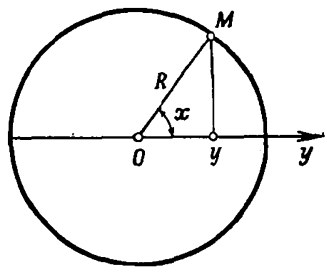


Fig. 1.1

(3) It sometimes turns out to be necessary in physics to investigate the behaviour of the function $y = f(x)$ whose argument x is itself a function $x = \varphi(t)$ of the new argument t. In that case it is said that y is a *composite function of the argument* t and x is referred to as an *intermediate* argument. This composite function can be represented by a single formula $y = f[\varphi(t)]$. Let us consider an example. Postulate that a particle M is rotating uniformly, with an angular velocity ω, along a circle of radius R. It is required to find the law governing the motion of the projection y of that particle on a certain axis Oy lying in the plane of the circle and passing through its centre O. We will assume that at time moment $t = 0$ the particle M is located on the axis Oy (Fig. 1.1). Let us denote by y the coordinate of the indicated projection on the axis Oy, and by x, the angle MOy. Evidently, $y = R \cos x$. On the other hand, since the particle moves along the circle with the angular velocity ω and at the time moment $t = 0$ is on the axis Oy, it follows that $x = \omega t$. Hence, y is a *composite function of the argument* t: $y = R \cos x$, where $x = \omega t$, or $y = R \cos \omega t$. Note that in mechanics the motion according to the law $y = R \cos \omega t$ is called *harmonic oscillation*.

1.2. INSTANTANEOUS VELOCITY AND THE NEW MATHEMATICAL CONCEPTS RELATED TO IT

1.2.1. Let the function $y = f(x)$ represent the law of motion of a particle along the Oy axis. To characterize the motion, it is important to know its average velocity. We shall now calculate the average velocity v_{av} of the moving particle during the time interval from x to $x + \Delta x$, where x is a fixed moment of time and Δx is a time

* The methods of representing a function will be given in more detail in Chap. 4.

increment. Since at the moment x the moving particle is at a distance $f(x)$ from the origin, and at the moment $x + \Delta x$ at a distance $f(x + \Delta x)$, the path Δy traversed by the particle during the time Δx is equal to $\Delta y = f(x + \Delta x) - f(x)$. Consequently, the average velocity is equal to

$$v_{av} = \frac{\Delta y}{\Delta x} = \frac{f(x+\Delta x)-f(x)}{\Delta x}.$$

Since the time moment x is fixed, it can be seen from the last formula that v_{av} is a function of the argument Δx. Besides the average velocity, a significant role in characterizing nonuniform motion is played by *instantaneous velocity* at the given moment of time x.

The *instantaneous velocity* (or simply *velocity*) at the time moment x is the number which the value of the average velocity

$$v_{av} = \frac{f(x+\Delta x)-f(x)}{\Delta x}$$

approaches when the time interval Δx tends to zero.

The physical concept of instantaneous velocity leads to the significant mathematical concept of a *derivative*. Abstracting from the concrete physical meaning of the function $y = f(x)$, we shall use the term *derivative* of this function at a fixed point x for the limit which the fraction $\frac{\Delta y}{\Delta x} = \frac{f(x+\Delta x)-f(x)}{\Delta x}$ approaches at Δx tending to zero.

The operation of finding a derivative is usually called *differentiation*. The derivative of the function $y = f(x)$ at a given fixed point is symbolized as $y'(x)$ or $f'(x)$. Using the accepted symbol for denoting the limit, we can write

$$f'(x) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x+\Delta x)-f(x)}{\Delta x}.$$

Let us consider some examples.

(1) Calculate the instantaneous velocity of a particle falling under the force of gravity. Since the law of motion of this particle is determined by the function $S = \frac{gt^2}{2}$, the path ΔS traversed by the particle during the time interval from t to $t + \Delta t$ is equal to

$$\Delta S = \frac{g(t+\Delta t)^2}{2} - \frac{gt^2}{2} = gt\,\Delta t + \frac{g}{2}(\Delta t)^2.$$

Therefore, the average velocity over the same time interval is equal to

$$v_{av} = \frac{\Delta S}{\Delta t} = gt + \frac{g}{2}\Delta t.$$

Consequently, the instantaneous velocity v at the fixed time moment t is equal to

$$v = \lim_{\Delta t \to 0} \frac{\Delta S}{\Delta t} = \lim_{\Delta t \to 0} \left(gt + \frac{g}{2} \Delta t \right) = gt.$$

We have actually calculated the derivative of the function $S = \frac{gt^2}{2}$, and so we can write $S' = gt$.

(2) Calculate the derivative of the function $y = x^n$, where n is a positive integer. Taking a fixed x and assuming Δx to be arbitrary, we obtain, by making use of Newton's binomial formula

$$\Delta y = (x + \Delta x)^n - x^n = nx^{n-1} \Delta x + \frac{n(n-1)}{2} x^{n-2} (\Delta x)^2 + \ldots$$

$$\ldots + (\Delta x)^n.$$

And so we see that the average velocity $\Delta y / \Delta x$ of variation of the function $y = f(x)$ in the interval from x to $x + \Delta x$ is equal to

$$\frac{\Delta y}{\Delta x} = nx^{n-1} + \frac{n(n-1)}{2} x^{n-2} \Delta x + \ldots + (\Delta x)^{n-1}.$$

Thus, the derivative at the given fixed point x is equal to

$$y' = \lim_{\Delta x \to 0} \left[nx^{n-1} + \frac{n(n-1)}{2} \Delta x + \ldots + (\Delta x)^{n-1} \right] = nx^{n-1}.$$

So we see that the concept of the limit of a function plays a very significant role in the calculation of a derivative. The elucidation of this concept is connected with the necessity to define the very concept of a function, of a variable quantity and a real number in more detail.

1.2.2. We shall now see that new mathematical problems arise in the process of calculating the derivatives of basic elementary functions.

We shall compute the derivative of the function $y = \sin x$. Taking a fixed value of x and an arbitrary Δx, we obtain

$$\Delta y = \sin(x + \Delta x) - \sin x = 2 \cos \left(x + \frac{\Delta x}{2} \right) \sin \frac{\Delta x}{2}.$$

Hence it follows that

$$\frac{\Delta y}{\Delta x} = \cos \left(x + \frac{\Delta x}{2} \right) \frac{\sin \left(\frac{\Delta x}{2} \right)}{\left(\frac{\Delta x}{2} \right)}.$$

Thus, to calculate the derivative of the function $y = \sin x$ at point x, it is necessary to find the following limit:

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \left[\cos \left(x + \frac{\Delta x}{2} \right) \frac{\sin \left(\frac{\Delta x}{2} \right)}{\left(\frac{\Delta x}{2} \right)} \right]. \qquad (1.1$$

It is natural to expect that for a fixed value of x

$$\lim_{\Delta x \to 0} \cos\left(x + \frac{\Delta x}{2}\right) = \cos x. \tag{1.2}$$

However, not every function $y = f(x)$ has the property

$$\lim_{\Delta x \to 0} f\left(x + \frac{\Delta x}{2}\right) = f(x).$$

Actually, this property means that when the argument of the function tends to the number x, the corresponding value of this function tends to the number $f(x)$. Functions possessing this property are called *continuous* (at point x). *Continuity* of a function is one of the most important concepts in mathematics.

To calculate the limit (1.1), it is necessary to compute, besides the limit (1.2), the limit

$$\lim_{\Delta x \to 0} \frac{\sin\left(\frac{\Delta x}{2}\right)}{\left(\frac{\Delta x}{2}\right)}. \tag{1.3}$$

This limit plays a significant role in mathematical analysis. It is often called the *first remarkable limit*. It can be proved that this limit is equal to unity and, therefore, the limit (1.1) is equal to $\cos x$.

Thus, we have

$$(\sin x)' = \cos x.$$

Let us take another example and calculate the derivative of the function $y = \log_a x$. Fixing the positive value of x and taking an arbitrary Δx (such that $x + \Delta x > 0$), we obtain

$$\Delta y = \log_a (x + \Delta x) - \log_a x = \log_a \left(1 + \frac{\Delta x}{x}\right),$$

whence

$$\frac{\Delta y}{\Delta x} = \frac{1}{\Delta x} \log_a \left(1 + \frac{\Delta x}{x}\right) =$$

$$= \frac{1}{x}\left[\frac{x}{\Delta x} \log_a \left(1 + \frac{\Delta x}{x}\right)\right] = \frac{1}{x} \log_a \left[\left(1 + \frac{\Delta x}{x}\right)^{\frac{x}{\Delta x}}\right].$$

From this it follows that to compute the derivative of the function $y = \log_a x$ at point x, we must find the limit

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{1}{x} \log_a \left[\left(1 + \frac{\Delta x}{x}\right)^{\frac{x}{\Delta x}}\right]. \tag{1.4}$$

Let us consider the limit of the expression in square brackets for $\Delta x \to 0$. It reduces to

$$\lim_{h \to 0} [(1+h)^{\frac{1}{h}}] \quad \left(\text{for } h = \frac{\Delta x}{x}\right).$$

This limit also plays an important role in the analysis and is often called the *second remarkable limit*. The existence of this limit can be easily proved. As it was first done by Euler*, the number equal to that limit is denoted by the letter e**, i.e.

$$\lim_{h \to 0} [(1+h)^{\frac{1}{h}}] = e. \tag{1.5}$$

Let us proceed with computing the limit (1.4). The argument of the logarithm in formula (1.4) is the quantity $\left[\left(1+\frac{\Delta x}{x}\right)^{\frac{x}{\Delta x}}\right]$, which, according to (1.5), tends to e when $\Delta x \to 0$. If a logarithmic function is continuous, then $\log_a \left[\left(1+\frac{\Delta x}{x}\right)^{\frac{x}{\Delta x}}\right]$ tends to $\log_a e$ for $\Delta x \to 0$. Thus, to determine the limit (1.4), we have to substantiate the continuity of the logarithmic function and make use of the limit (1.5). Assuming that it has already been done, we obtain the value of the limit (1.4) which is equal to $\frac{1}{x}\log_a e$. Thus,

$$(\log_a x)' = \frac{1}{x}\log_a e.$$

We shall not compute here the derivatives of other basic elementary functions such as $y = \cos x$, $y = \tan x$, $y = \cot x$, $y = \arcsin x$, $y = \arccos x$, $y = \arctan x$, $y = \operatorname{arccot} x$, $y = a^x$ and $y = x^\alpha$, where α is an arbitrary number. In computations of the derivatives of these functions no other difficulties are encountered besides those indicated above. Indeed, to compute the derivatives of all basic elementary functions, only their continuity and the two remarkable limits will be needed.

Given below are the derivatives of the basic elementary functions.

1°. $(x^\alpha)' = \alpha x^{\alpha-1}$, α being an arbitrary number.

2°. $(\log_a x)' = \frac{1}{x}\log_a e$, in particular, if $a = e$, then $(\log_e x)' = \frac{1}{x}$.

3°. $(a^x)' = a^x \log_e a$, in particular, if $a = e$, then $(e^x)' = e^x$.

---

* Leonhard Euler (1707-1783), an eminent mathematician, born in Switzerland, most of his life lived in Russia, was a Member of the St. Petersburg Academy of Sciences.

** In Sec. 8.16 we shall indicate the method of calculating the number e with any degree of accuracy. We shall also give there the result of calculating the number e by a computer, with an accuracy to within 590 decimal places.

4°. $(\sin x)' = \cos x$.
5°. $(\cos x)' = -\sin x$.
6°. $(\tan x)' = \dfrac{1}{\cos^2 x}$.
7°. $(\cot x)' = -\dfrac{1}{\sin^2 x}$.
8°. $(\arcsin x)' = \dfrac{1}{\sqrt{1-x^2}}$.
9°. $(\arccos x)' = -\dfrac{1}{\sqrt{1-x^2}}$.
10°. $(\arctan x)' = \dfrac{1}{1+x^2}$.
11°. $(\operatorname{arc\,cot} x)' = -\dfrac{1}{1+x^2}$.

1.2.3. To compute the derivatives of a variety of functions in addition to the table given above use should be made of *the rule of differentiation of a composite function*, as well as *the rules of differentiation of the sum, the difference, the product and the quotient of the functions.*

The rule of differentiating the composite function $y = f(x)$, where $x = \varphi(t)$ reads as follows.

To determine the derivative $y'(t)$ of the composite function $y = f[\varphi(t)]$ with respect to its argument t at a given point t, it is necessary (1) *to compute the derivative $\varphi'(t)$ of the function $x = \varphi(t)$ at point t*; (2) *to compute the derivative $f'(x)$ of the function $y = f(x)$ at point x, where $x = \varphi(t)$*; (3) *to multiply the indicated derivatives.* Thus, the derivative of the composite function $y = f[\varphi(t)]$ can be found from the formula $y'(t) = f'(x)\,\varphi'(t)$. The reasoning that follows will explain the formulated rule. Let us give an arbitrary increment $\Delta t \neq 0$ to the argument t at point t. To this increment there corresponds the increment $\Delta x = \varphi(t + \Delta t) - \varphi(t)$ of the function $x = \varphi(t)$. To the resultant increment Δx there corresponds the increment $\Delta y = f(x + \Delta x) - f(x)$ of the function $y = f(x)$ at point x. Neglecting the case $\Delta x = 0$, we shall consider the relation

$$\frac{\Delta y}{\Delta t} = \frac{\Delta y}{\Delta x}\,\frac{\Delta x}{\Delta t}.$$

Since $\lim\limits_{\Delta t \to 0} \dfrac{\Delta x}{\Delta t} = \varphi'(t)$ and $\lim\limits_{\Delta x \to 0} \dfrac{\Delta y}{\Delta x} = f'(x)$, and it is evident from the existence of the first of these limits that $\Delta x \to 0$ when $\Delta t \to 0$,* it follows that the limit $\lim\limits_{\Delta t \to 0} \dfrac{\Delta y}{\Delta t}$ exists and is equal to $f'(x)\,\varphi'(t)$, i.e. $y'(t) = f'(x)\,\varphi'(t)$.

Now we shall give the rules of differentiation of the sum, the difference, the product and the quotient (under the assumption that

* If the denominator of the fraction having a limit tends to zero, then the numerator of that fraction also tends to zero.

$u(x)$ and $v(x)$ have derivatives):

$$[u(x) \pm v(x)]' = u'(x) \pm v'(x),$$

$$[u(x)\, v(x)]' = u(x)\, v'(x) + u'(x)\, v(x),$$

$$\left[\frac{u(x)}{v(x)}\right]' = \frac{u'(x)\, v(x) - u(x)\, v'(x)}{v^2(x)}.$$

By way of an example, we will demonstrate the derivation of the second of these formulas. We shall impart an arbitrary increment $\Delta x \neq 0$ to the argument x. To this increment Δx there corresponds the increment Δy of the function $y = u(x)\, v(x)$:

$$\begin{aligned}\Delta y &= u(x + \Delta x)\, v(x + \Delta x) - u(x)\, v(x) = \\ &= u(x + \Delta x)\, [v(x + \Delta x) - v(x)] + \\ &+ v(x)\, [u(x + \Delta x) - u(x)] = u(x + \Delta x)\, \Delta v + v(x)\, \Delta u.\end{aligned}$$

Thus we have

$$\frac{\Delta y}{\Delta x} = u(x + \Delta x)\, \frac{\Delta v}{\Delta x} + v(x)\, \frac{\Delta u}{\Delta x}.$$

Since there exist the limits $\lim\limits_{\Delta x \to 0} \frac{\Delta u}{\Delta x} = u'(x)$ and $\lim\limits_{\Delta x \to 0} \frac{\Delta v}{\Delta x} = v'(x)$, and it is clear from the existence of the first of these limits that $\lim\limits_{\Delta x \to 0} u(x + \Delta x) = u(x)$, it follows that the limit $\lim\limits_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$ exists and is equal to $u(x)\, v'(x) + v(x)\, u'(x)$.

Here are some examples on the application of the indicated rules:

(1) Calculate the derivative of the function $y = cu(x)$, where c is a constant. It is easy to verify that the derivative of a constant is equal to zero. Consequently, by the formula of differentiation of a product we obtain $[cu(x)]' = cu'(x)$.

(2) Calculate the derivative of the function $y = \tan x$. Since $\tan x = \frac{\sin x}{\cos x}$, the differentiation formula for the quotient yields

$$(\tan x)' = \frac{(\sin x)' \cos x - \sin x\, (\cos x)'}{\cos^2 x} = \frac{1}{\cos^2 x}.$$

(3) Calculate the derivative of the function describing the harmonic oscillations, $y = A \cos(\omega t + \delta)$, where A, ω and δ are constants. We shall regard this function as a composite function of the form $y = A \cos x$, where $x = \omega t + \delta$. According to the rule of differentiation of a composite function we obtain

$$y'(t) = (A \cos x)'\, (\omega t + \delta)' = -(A \sin x)\, \omega,$$

where $x = \omega t + \delta$. Therefore,

$$y'(t) = -A\omega \sin(\omega t + \delta).$$

(4) Calculate the derivative of the function $y = a^{\arctan t}$. We shall consider this function to be a composite function of the form $y = a^x$, where $x = \arctan t$. According to the rule of differentiation of a composite function

$$y'(t) = (a^x)'(\arctan t)' = (a^x \log_e a)\left(\frac{1}{1+t^2}\right),$$

where $x = \arctan t$. Therefore

$$(a^{\arctan t})' = \frac{a^{\arctan t} \log_e a}{1+t^2}.$$

The differentiation rules and the table of derivatives given above are the principal devices used in the part of mathematical analysis which is usually called *differential calculus*. Thus, one of the main

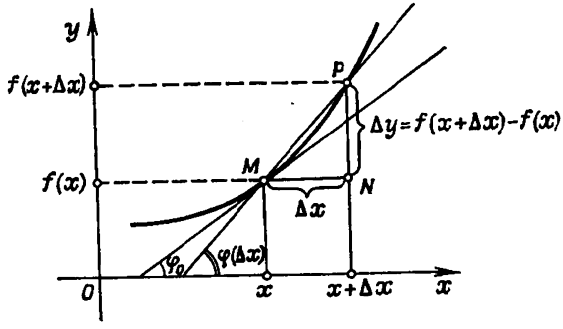


Fig. 1.2

tasks of differential calculus is the substantiation of all the formulas given in the table of derivatives and the rules of differentiation of the sum, the difference, the product and the quotient, and of a composite function.

1.2.4. Let us elucidate the geometrical meaning of a derivative. For this purpose we shall consider the graph of the function $y = f(x)$* (Fig. 1.2). Let a point M on the graph of the function correspond to the fixed value of the argument x, and a point P, to the value $x + \Delta x$, where Δx is an increment of the argument. The line MP will be called a *secant*. Let us designate by $\varphi(\Delta x)$ the angle made by the secant with the axis Ox (this angle is, evidently, dependent on Δx). *The tangent to the graph of the function* $y = f(x)$ *at point* M *will be taken to be the limiting position occupied by the secant* MP *when point* P *tends to point* M *along the graph* (or, *what is the same*,

* The graph of the function $y = f(x)$ is the locus of the points of the plane, for each of which the ordinate is the value y of this function corresponding to the abscissa x.

when Δx *tends to zero*). It is clear from Fig. 1.2 that

$$\tan \varphi (\Delta x) = \frac{PN}{MN} = \frac{\Delta y}{\Delta x} = \frac{f(x+\Delta x) - f(x)}{\Delta x}.$$

Since for $\Delta x \to 0$ the secant MP turns into a tangent, we have

$$\lim_{\Delta x \to 0} \tan \varphi (\Delta x) = \tan \varphi_0,$$

where φ_0 is the angle formed by the tangent with the Ox axis. On the other hand,

$$\lim_{\Delta x \to 0} \tan \varphi (\Delta x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} = f'(x).$$

Consequently, $f'(x) = \tan \varphi_0$. The tangent of the angle that the line makes with the x-axis is called the *slope* of that line. Thus *the derivative* $f'(x)$ *is equal to the slope of the tangent to the graph of the function* $y = f(x)$ *at point* M.

1.3. THE PROBLEM OF RECONSTRUCTING THE LAW OF MOTION FROM THE VELOCITY AND THE RELATED MATHEMATICAL PROBLEMS

Let us consider the following problem from physics. Assume that the instantaneous velocity $f(x)$ of a particle moving along the Oy axis is given for any moment of time x and the position y_0 of that particle at the initial moment of time $x = x_0$ is known. The problem is to find the law of motion of that particle.

Since the instantaneous velocity $f(x)$ is the derivative of the function $y = F(x)$ specifying the law of motion of the particle along the y-axis, the problem reduces to determining, from a given function $f(x)$, a function $F(x)$ whose derivative $F'(x)$ is equal to $f(x)$.

Abstracting from the concrete physical meaning of the functions $f(x)$ and $F(x)$, we shall arrive at the mathematical notions of *antiderivative* and *indefinite integral. An antiderivative of the function* $f(x)$ *is such a function* $F(x)$ *whose derivative* $F'(x)$ *is equal to* $f(x)$.

It is evident that if the function $F(x)$ is an antiderivative of the function $f(x)$, then the function $F(x) + C$, where C is an arbitrary constant, is, in its turn, the antiderivative of the function $f(x)$ (since the derivative of the constant C is zero).

It can be proved that any two antiderivatives of the same function $f(x)$ differ by a constant value. Thus, if the function $F(x)$ is one of the antiderivatives of the function $f(x)$, then *any antiderivative of the function* $f(x)$ *has the form* $F(x) + C$, where C is a constant.

The totality of all antiderivatives of the same function $f(x)$ *is called the indefinite integral of the function* $f(x)$ *and is symbolized as* $\int f(x)\,dx$.

Consequently, if $F(x)$ is one of the antiderivatives of the function $f(x)$, then

$$\int f(x)\,dx = F(x) + C.$$

Now let us return to the solution of the physical problem posed above. The law of motion of a particle having an instantaneous velocity $f(x)$, which we are interested in, is defined by the function $y = F(x) + C$, where $F(x)$ is an antiderivative of the function $f(x)$, C being a certain constant. To determine the constant C, we shall make use of the fact that $y = y_0$ at the initial moment of time $x = x_0$, that is $y_0 = F(x_0) + C$, whence it follows that $C = y_0 - F(x_0)$. Hence, the law of motion we wish to determine has the form

$$y = F(x) + y_0 - F(x_0).$$

We shall now consider some examples from physics and mathematics.

(1) The instantaneous velocity of a particle moving along the y-axis has the form $f(x) = \cos x$. It is required to find the law of motion of that particle if at the initial moment of time $x = x_0$ the particle is at the point $y = y_0$ of the Oy axis. It is clear from the table of the derivatives that one of the antiderivatives of the function $f(x) = \cos x$ is the function $F(x) = \sin x$. Consequently, the required law of motion has the form $y = \sin x + C$.

From the condition $y = y_0$, with $x = x_0$, we find $C = y_0 - \sin x_0$, that is, the final form of the law of motion is

$$y = \sin x + y_0 - \sin x_0.$$

(2) It is required to find $\int \frac{1}{1+x^2}\,dx$. It is clear from the table of the derivatives that one of the antiderivatives of the function $f(x) = \frac{1}{1+x^2}$ is the function $F(x) = \arctan x$. Hence

$$\int \frac{1}{1+x^2}\,dx = \text{arc tan}\, x + C.$$

In the previous section we presented a table of derivatives of elementary functions. Taking into account that every formula $F'(x) = f(x)$ from that table leads to the corresponding formula $\int f(x)\,dx = F(x) + C$, we obtain the following table of indefinite integrals:

1°. $\int x^\alpha\,dx = \frac{x^{\alpha+1}}{\alpha+1} + C \quad (\alpha \neq -1)$.

2°. $\int \frac{dx}{x} = \log_e |x| + C$.

3°. $\int a^x\,dx = \frac{a^x}{\log_e a} + C$.

$4°.\ \int \sin x\,dx = -\cos x + C.$

$5°.\ \int \cos x\,dx = \sin x + C.$

$6°.\ \int \frac{dx}{\cos^2 x} = \tan x + C.$

$7°.\ \int \frac{dx}{\sin^2 x} = -\cot x + C.$

$8°.\ \int \frac{dx}{\sqrt{1-x^2}} = \arcsin x + C.$

$9°.\ \int \frac{dx}{1+x^2} = \arctan x + C.$

This table of indefinite integrals, together with integration rules (not given here), constitutes an important apparatus used in the part of the analysis usually called *integral calculus.*

However, this apparatus turns out to be insufficient to compute a considerable number of indefinite integrals. A problem arises concerning the existence of an antiderivative (and an indefinite integral) of an arbitrary function $f(x)$, continuous at every point x. In the next section we shall consider another approach to the problem of integrating a function which will make it possible to solve the problem.

For the time being, it suffices to point out that there are continuous (at each point x) functions ($y = \cos x^2$, for example), whose antiderivatives exist but cannot be represented by means of a finite number of operations of addition, subtraction, multiplication, division and the formation of composite functions from the basic elementary functions given in 1.2.2.

1.4. SOME QUESTIONS ARISING IN SOLVING THE PROBLEMS ON COMPUTATION OF A PATH

1.4.1. Let the function $f(x)$ represent the velocity of a particle moving along the y-axis. For the sake of simplicity we shall assume all the values of the function $f(x)$ to be nonnegative. It is required to calculate the path traversed by the particle during the time interval from $x = a$ to $x = b$.

To solve this problem* we shall partition the time interval under consideration into small subintervals bounded by the time moments $a = x_0 < x_1 < x_2 < \ldots < x_n = b$. It is natural to assume that the velocity $f(x)$ changes but very slightly in each of the subintervals between x_{k-1} and x_k. Therefore, we can consider the velocity in the indicated interval to be approximately constant and equal, say, to $f(x_k)$. In that case, the path traversed by the particle during the time

* The connection between this problem and that considered in the previous section will be shown later on.

interval $\Delta x_k = x_k - x_{k-1}$ is approximately equal to $f(x_k)\,\Delta x_k$, and the path S_a^b traversed by the particle during the time interval between a and b is approximately equal to

$$S_a^b \approx f(x_1)\,\Delta x_1 + f(x_2)\,\Delta x_2 + \ldots + f(x_n)\,\Delta x_n. \tag{1.6}$$

It is natural to expect that making the time intervals Δx_k smaller and smaller we shall obtain the more and more accurate value of the path S_a^b. We shall have the most accurate value of the path S_a^b if we

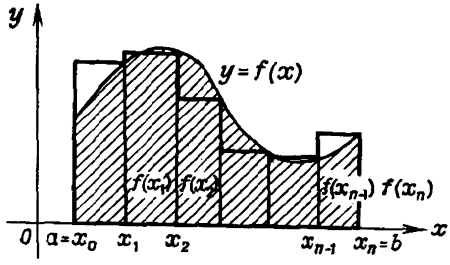


Fig. 1.3

pass to the limit in the sum (1.6), with all Δx_k tending to zero (in that case the number of summands in (1.6) will be sure to increase indefinitely). Using the symbol of the limit, we shall be able to write

$$S_a^b = \lim_{\Delta x_k \to 0} [f(x_1)\,\Delta x_1 + f(x_2)\,\Delta x_2 + \ldots + f(x_n)\,\Delta x_n]. \tag{1.7}$$

Since we must explain what we mean by the limit of the sum written above, we have of necessity to consider the concept of the limit in more detail. In mathematics, the limit (1.7) is called the *definite integral* of the function $f(x)$ in the limits from a to b and is denoted by the symbol

$$S_a^b = \int_a^b f(x)\,dx.$$

The sum (1.6) is the sum of the areas of the rectangles whose bases are the segments Δx_k and whose altitudes are $f(x_k)$. In other words, this sum is equal to the area of the stepped figure illustrated in Fig. 1.3 (the stepped figure is outlined by a bold line). It is natural to expect that with the lengths of all the segments Δx_k tending to zero, the area of the indicated stepped figure will tend to the area of the hatched curvilinear figure lying below the graph of the function $y = f(x)$ in the interval between a and b. This curvilinear figure is often called a *curvilinear trapezoid*. Thus the definite integral is equal to the area of the indicated curvilinear trapezoid.

This was a preliminary discussion, and now we shall elucidate the notion of the area of a curvilinear trapezoid and, in general, the area of a plane figure.

1.4.2. We see that two important problems are connected with the notion of the definite integral: the physical problem on the computation of a path and a geometrical problem on the computation of the area of a plane figure. Consequently, the consideration of the means of computing the definite integral is very important indeed.

Let us denote by $F(x)$ the definite integral of the function $f(x)$ in the limits between a and x, i.e. put

$$F(x) = \int_a^x f(x)\,dx.$$

From the geometrical point of view, the integral is equal to the area of the curvilinear trapezoid located below the graph of the function $y = f(x)$ in the interval between a and x. In Fig. 1.4 this

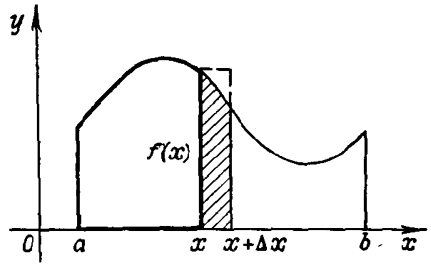


Fig. 1.4

trapezoid is drawn in bold line. Reasoning in terms of geometry, we shall show that the function $F(x)$ that has been introduced is one of the antiderivatives of the function $f(x)$, i.e. we shall make sure that $F'(x) = f(x)$. Let Δx be a certain increment of the argument x. Evidently, the difference $F(x + \Delta x) - F(x)$ is equal to the area of the "narrow" curvilinear trapezoid hatched in Fig. 1.4. When Δx is small, the area of this trapezoid differs little from the area $f(x)\Delta x$ of the rectangle with the base Δx and the altitude $f(x)$. It follows that for small Δx the ratio

$$\frac{F(x+\Delta x) - F(x)}{\Delta x} \tag{1.8}$$

differs little from the altitude $f(x)$ of the rectangle indicated above. Since, at $\Delta x \to 0$, the limit of the fraction (1.8) is equal to the derivative $F'(x)$, we have $F'(x) = f(x)$. Thus, the function $F(x)$ is one of the antiderivatives of the function $f(x)$. Consequently, *any* antiderivative $\Phi(x)$ of the function $f(x)$ has the form

$$\Phi(x) = F(x) + C = \int_a^x f(x)\,dx + C. \tag{1.9}$$

Of course, our reasoning and the resulting formula (1.9) are valid, in a general case, not for every function $f(x)$. It is easy to prove the validity of formula (1.9) for any function $f(x)$ continuous at every point x. Thus, the determination of formula (1.9) solves the problem of the existence of an antiderivative (and an indefinite integral) of any function $f(x)$ continuous at every point x.

Let us now determine, with the aid of the same formula (1.9), the connection between the definite integral $\int_a^b f(x)\,dx$ and any antiderivative $\Phi(x)$ of the function $f(x)$. Putting in succession $x = a$ and $x = b$ in formula (1.9) and taking into account the obvious (in terms of geometry) equality

$$\int_a^a f(x)\,dx = 0,$$

we obtain

$$\Phi(a) = \int_a^a f(x)\,dx + C = C, \quad \Phi(b) = \int_a^b f(x)\,dx + C.$$

Therefore,

$$\int_a^b f(x)\,dx = \Phi(b) - \Phi(a). \tag{1.10}$$

Formula (1.10) is one of the basic formulas of integral calculus and is known as *Newton-Leibniz** formula. This formula reduces the problem of calculating a definite integral to that of computing an antiderivative (or an indefinite integral). The substantiation of this formula is one of the most important tasks of mathematical analysis. A number of methods exist for an approximate calculation of definite integrals, the simplest of which is based on the replacement of that integral by the sum (1.6). These methods and relation (1.9) make possible an approximate calculation of indefinite integrals as well; in particular, they enable us to calculate the antiderivative of any continuous (at every point x) function $f(x)$.

Let us compute, for example, the area S_1 bounded by the graph of the function $y = \sin x$ in the interval between 0 and π and the x-axis

* Gottfried Wilhelm von Leibniz, German philosopher and mathematician (1646-1716).

(Fig. 1.5)*. Proceeding from the assertions made above, we can write

$$S_1 = \int_0^{\pi} \sin x \, dx.$$

Since the function $\Phi(x) = -\cos x$ is one of the antiderivatives of the function $f(x) = \sin x$, we obtain from formula (1.10)

$$S_1 = \int_0^{\pi} \sin x \, dx = (-\cos \pi) - (-\cos 0) = 2.$$

Let us now calculate the area S_2 of the figure intercepted from the parabola $y = x^2$ by a straight line passing through two points

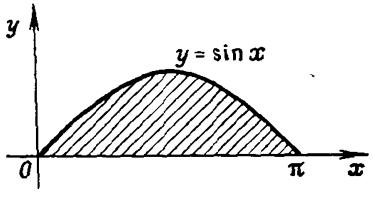


Fig. 1.5

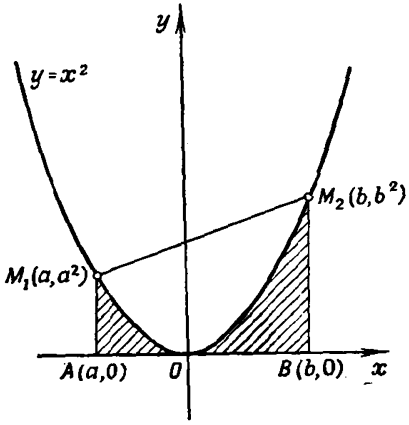


Fig. 1.6

$M_1(a, a^2)$ and $M_2(b, b^2)$ of this parabola (Fig. 1.6)**. The area S_2 sought is equal to the difference between the areas of the rectilinear trapezoid AM_1M_2B and the curvilinear trapezoid hatched in the figure, i.e.

$$S_2 = \frac{(b^2+a^2)(b-a)}{2} - \int_a^b x^2 \, dx = \frac{(b^2+a^2)(b-a)}{2} - \frac{b^3-a^3}{3} = \frac{(b-a)^3}{6}.$$

1.5. CONCLUDING REMARKS

Differential and integral calculus forms the basis of mathematical analysis, which is one of the greatest achievements of human mind. The introduction into mathematics of the concepts of a variable

* It is rather difficult to calculate this area by the means of elementary mathematics.

** This problem was solved by Archimedes, the eminent scientist of ancient Greece, in the 3rd century B. C. by the means of elementary mathematics.

quantity and a function made it possible to pass from the solution of various individual physical and geometrical problems to the creation of general methods for solving these problems. The development of differential and integral calculus had a tremendous effect on the general progress of science and technology.

We believe that any further progress in science and technology can only be possible if all our conceptions of nature are reconsidered in terms of mathematics, and this, in turn, requires the development of new branches of mathematics. We can say for sure that the mathematization of our conceptions, strict quantitative formulation of the regularities and the wide use of computers and computational technique constitute the basis of the development of natural sciences.

The use of modern computational technique significantly simplifies, as a rule, the most labour-intensive and complicated calculations*. But at the same time a lot of mathematical problems arise, such as the working out of algorithms** of computations which can serve as the source of compiling programs for computers, the development of the control theory, the theory of optimal processes, mathematical logic and theoretical cybernetics.

In the following chapters we shall deal with the construction of the apparatus of mathematical analysis. We shall also consider some applications of this apparatus to the development of numerical algorithms.

Having considered the preliminary information, we have now before us the following problems.

1. The elucidation of the concepts of a real number, a variable quantity and a function.
2. The definition and development of the concept of the limit of a function and the related notion of the continuity of a function.
3. The substantiation of the formulas and rules of the differential and integral calculus.
4. The construction of the theory of definite integral as the limit of the sums of a particular kind and the development of the methods of calculating the definite integral.
5. Elucidation of some geometrical notions (the area of a plane figure, the length of an arc, etc.).

---

* It takes modern computers several minutes to carry out calculations on which a human being would spend the whole of his life.

** Algorithm is a system of computations carried out according to strict rules and leading, after a number of steps, to the solution of the problem.

CHAPTER 2

THEORY OF REAL NUMBERS

From the elementary course the reader is sure to have the notion of real numbers and of their necessity, for instance, for measuring line segments and intervals of time. To have a deeper knowledge of the most important mathematical concepts, such as a variable quantity, a function and a limit, it is necessary to develop the theory of real numbers still further.

Consider, for example, such a variable physical quantity as time. To make a comparison between various time intervals, we have to know how to compare real numbers. In other words, we should establish a rule enabling us to decide which of the two given real numbers is greater. The practice of successive measurements of time leads to the necessity of defining the operations of addition and multiplication of real numbers and elucidating the properties of these operations. It is also necessary to establish the properties of real numbers in order to substantiate the applicability of the rules of elementary algebra to these numbers.

2.1. REAL NUMBERS

2.1.1. The properties of rational numbers. Recall that a rational number is a number which can be represented as a ratio of two integers.* Addition and multiplication of rational numbers, the rules of comparison of these numbers and their simplest properties are sure to be known from the elementary course. Here we shall speak of the main properties of rational numbers following from the corresponding properties of integral numbers.

The following three rules are the basic properties of rational numbers: **the rule of comparison and the rules of finding a sum and a product.**

I. *Any two rational numbers a and b are connected by one and only one of the three signs:* $>$, $<$ *or* $=$, *and if* $a > b$, *then* $b < a$. In other words, *there exists a rule making it possible to determine which of the*

* One and the same rational number can be represented as the ratios of different integers, for instance, $1/2 = 2/4 = 3/6 = \ldots$.

three indicated signs connects two given rational numbers. This rule is called the **rule of comparison.***

II. *There is a rule according to which any two rational numbers a and b are associated with a definite rational number c called their sum and symbolized as* $c = a + b$.**

The operation of finding a sum is called **summation.**

III. *There exists a rule according to which any two rational numbers a and b are associated with a definite rational number c called their product and symbolized as* $c = ab$.*** The operation of finding a product is called **multiplication.**

We shall now enumerate the basic properties which govern the three indicated rules.

The rule of comparison of rational numbers has the following properties.

1°. *It follows from* $a > b$ *and* $b > c$ *that* $a > c$ (*the transitivity of the sign* $>$); *it follows from* $a = b$ *and* $b = c$ *that* $a = c$ (*the transitivity of the sign* $=$).

The rule of addition of rational numbers has the following properties:

2°. $a + b = b + a$ (commutativity).

3°. $(a + b) + c = a + (b + c)$ (associativity).

4°. *There is a rational number* 0 *such that* $a + 0 = a$ *for any rational number a* (particular role played by zero).

5°. *For every rational number a there exists an inverse number* a' *such that* $a + a' = 0$.

The rule of multiplication of rational numbers has the following properties:

6°. $ab = ba$ (commutativity).

7°. $(ab)\,c = a\,(bc)$ (associativity).

8°. *There exists a rational number unity such that* $a \cdot 1 = a$ *for any rational number a* (a particular role played by unity).

9°. *For every rational number a, different from zero, there is an inverse number* a' *such that* $aa' = 1$.

The rules of addition and multiplication are connected by the following property:

---

* The rule of comparison of rational numbers is formulated as follows: two nonnegative rational numbers $a = m_1/n_1$ and $b = m_2/n_2$ are connected by the same sign as the two integers m_1n_2 and m_2n_1; two nonpositive rational numbers a and b are connected by the same sign as the two nonnegative numbers $|b|$ and $|a|$; if a is nonnegative and b is a negative rational number, then $a > b$.

** The rule of summation of the rational numbers $a = m_1/n_1$ and $b = m_2/n_2$ is determined by the formula $\frac{m_1}{n_1} + \frac{m_2}{n_2} = \frac{m_1n_2 + m_2n_1}{n_1n_2}$.

*** The rule of multiplication of rational numbers is determined by the formula $\frac{m_1}{n_1}\frac{m_2}{n_2} = \frac{m_1m_2}{n_1n_2}$.

10°. $(a + b)\,c = ac + bc$ (distributive property of multiplication with respect to the sum).

The following two properties relate the sign $>$ to the signs of addition and multiplication:

11°. *It follows from* $a > b$ *that* $a + c > b + c$.

12°. *It follows from* $a > b$ *and* $c > 0$ that $ac > bc$.

And here is the last property which plays a particular role.

13°. *Whatever the rational number* a, *the number* 1 *can be repeated as a summand as many times as is needed to obtain a sum exceeding* a.*

The thirteen properties enumerated above are usually called the *basic properties* of rational numbers, for all the other algebraic properties of these numbers, related to the arithmetic operations and to the combination of equalities and inequalities, can be derived from the indicated basic properties.

Thus, for instance, these properties yield, as their consequence, the property which will be often used later on, making possible the termwise summation of inequalities of like signs:

if $a > b$ and $c > d$, then $a + c > b + d$.

Indeed, it follows from the inequalities $a > b$ and $c > d$ and from properties 11° and 12° that $a + c > b + c$ and $b + c > b + d$, and from the last inequalities and property 1° it follows that $a + c > b + d$.

2.1.2. Measuring the segments of the number axis. It is known from the elementary course that two line segments can be *commensurable* (when the ratio of their lengths is a rational number) or *incommensurable* (the diagonal and a side of a square may serve as an example).

It is convenient at this time to introduce into consideration the number axis which is a straight line on which we have chosen a definite point O (an origin), a unit length OE (it is assumed to be equal to unity) and a positive direction (usually from O to E).

A question naturally arises as to the possibility of putting every point M of the number axis into correspondence with a certain number expressing the length of the segment OM. We shall assume this number to be positive if M and E are to the same side of O, and negative if otherwise.

Note first of all that *to every rational number there corresponds a definite point on the number axis.* Indeed, we know from the elementary course how to construct a line segment whose length is $1/n$-th of the unit length OE (n is any positive integer). Consequently, we can plot a segment AB whose length and the unit length OE are in the ratio $\frac{m}{n}$, where m and n are any positive integers. Assuming that point E is to the right of O (Fig. 2.1) and marking off the segment AB to the

* This property is often called the Archimedes axiom.

right (to the left) of point O, we shall obtain a point M_1 (M_2), corresponding to the rational number $+\frac{m}{n}\left(-\frac{m}{n}\right)$.

At the same time, the existence of incommensurable segments enables us to conclude that *not all the points of the number axis correspond to rational numbers.*

So the necessity naturally arises to expand the domain of rational numbers and to introduce into consideration the numbers which

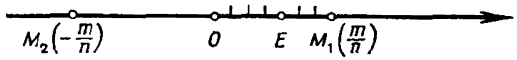


Fig. 2.1

Fig. 2.2

would correspond to all the numbers of the number axis and would make it possible to measure any line segment with the aid of the unit length OE.

We shall now describe the process of measuring the segment OM of the number axis and show that this process *puts any point M of that axis into correspondence with a certain quite definite infinite decimal fraction.*

Let M be any point of the number axis. For definiteness, we shall assume that M (as well as E) is to the right of O (Fig. 2.2).

We shall measure the line segment OM by means of the unit length OE. We shall find, first of all, how many times the whole segment OE can be laid off along the segment OM. Two cases are possible here:

(1) The segment OE can be laid off along the segment OM an integral number of times a_0 with a certain remainder NM smaller than OE (see Fig. 2.2). In this case the integer a_0 is a *minor* approximation of measuring with an accuracy to within unity.

(2) The segment OE can be laid off along the segment OM an integral number of times $a_0 + 1$ without a remainder. In this case, as well, the number a_0 is a *minor* approximation, with an accuracy to within unity, since the segment OE can be laid off along the segment OM a_0 times with the remainder NM equal to OE.*

* In practical measurements, the second process is usually considered to be completed and the length of the segment OM is taken to be $a_0 + 1$. However, it is convenient for us (for the sake of uniformity) to make a minor approximation, in order in this case, too, to obtain the remainder NM and be able to continue measurement.

Let us now find how many times one-tenth of the unit length OE can be laid off along the remainder NM. Here, again, two cases are possible.

(1) One-tenth of the segment OE can be contained in the segment NM an integral number of times a_1 with a certain remainder PM smaller than one-tenth of the segment OE (see Fig. 2.2). In this case the rational number a_0, a_1 is the result of the *minor* approximation with an accuracy to within one-tenth.

(2) One-tenth of the segment OE is contained in the segment NM an integral number of times $a_1 + 1$, this time without a remainder. In this case the rational number a_0, a_1 is again the result of the *minor* approximation with an accuracy to within one-tenth, since the one-tenth part of OE can be laid off along the segment NM a_1 times with a remainder PM equal to one-tenth of OE*.

Continuing our argument indefinitely, we shall come to an infinite set of rational numbers

$$a_0;\ a_0, a_1;\ \ldots;\ a_0, a_1a_2 \ldots a_n;\ \ldots, \tag{2.1}$$

each of which is the result of a *minor* approximation of the line segment OM with a corresponding degree of accuracy. At the same time, each of the numbers (2.1) can be obtained by terminating, at the appropriate decimal place, the *infinite decimal fraction*

$$a_0, a_1a_2 \ldots a_n \ldots . \tag{2.2}$$

The considerations carried out above are also applicable to the case when point M is to the left of point O, only in this case all the numbers (2.1) and the infinite decimal (2.2) will have a negative sign.

Thus we have established that *the above-described process of measuring the line segment OM can put into correspondence any point M of the number axis with a definite infinite decimal fraction.*

We see that this process of measuring an arbitrary segment OM of the number axis by means of a unit length leads us, in a natural way, to a consideration of *numbers representable as infinite decimal fractions.* At the same time, every infinite decimal (2.2) can be completely characterized by the infinite set (2.1) of rational numbers approximating the fraction. The process of measuring the segment OM described above can certainly be changed so that it will lead to the consideration of infinite binary fractions or of infinite fractions in any other number system.

Note that in modern computers a binary system is usually used to specify the numbers, and sometimes a ternary system. The reason is that radio valves or semiconductors used in computers usually have two, and sometimes three, stable states (for example, the valve is

* See the previous footnote.

non-conductive, no current flows—one stable state; the valve is conductive and a current flows—the other stable state; the third stable state is when the direction of the current flow is taken into consideration).

In view of this circumstance, a necessity arises of developing algorithms of converting numbers from the decimal notation to the binary one and back again. Examples of such algorithms can be found in Supplement 1 to the present chapter.

2.1.3. Real numbers and the rule of their comparison. *Let us consider a set of various infinite decimals. The numbers representable by these fractions will be called real.**

We shall call the given real number positive (negative) if it can be represented by a positive (negative) infinite decimal fraction.

The set of real numbers includes, naturally, all rational numbers since they are all representable by infinite decimals. The argument underlying the representation of a rational number by an infinite decimal is as follows. Any rational number is associated with a certain point M of the number axis, and this point, in its turn, can be associated, by the method indicated in Sec. 2.1.2, with a definite infinite decimal fraction. Thus, the rational number $\frac{1}{2}$ is associated with the infinite decimal fraction 0.4999 . . ., the rational number $\frac{4}{3}$, with the infinite decimal 1.333 . . ..

Real numbers other than rational are called *irrational.*

Our problem is to carry over, in succession, the three rules and all the *basic properties* of rational numbers, enumerated in Sec.2.1.1, to the case of arbitrary real numbers. In that way the applicability of all the rules of elementary algebra related to arithmetic operations and to combinations of equalities and inequalities will be substantiated for the case of real numbers.

In this subsection we shall establish the rule of comparing real numbers. Before formulating this rule, we shall define the notation that we shall accept for the rational numbers *representable as finite decimal fractions.* Note that these rational numbers can be written as infinite decimal fractions *in two ways.* For example, the number $\frac{1}{2} = 0.5$ can be written as $\frac{1}{2} = 0.4999\ldots$; and as $\frac{1}{2} = 0.5000\ldots$.

And, in general, the rational number $a_0, a_1a_2 \ldots a_n$, where $a_n \neq 0$, can be written in the form $a_0, a_1a_2 \ldots a_{n-1}(a_n - 1)\,999\ldots$; and in the form $a_0, a_1a_2 \ldots a_n\,000\ldots$.

The first of these two notations can be obtained by the method indicated in 2.1.2 and the second, by a formal transformation of the

* As was mentioned in the footnote on p 18, the concept of a number is one of the *basic concepts.*

given finite decimal into a continuous decimal by adding zeros to the notation.

When we compare real numbers, we shall agree to use the first of these notations in the form of an infinite decimal for the indicated rational numbers.

In other words, when comparing real numbers we shall not employ infinite decimal fractions whose all decimal digits, beginning with a certain decimal place, are equal to zero (except for the fraction 0.000, of course).*

Now we shall formulate the *rule of comparison* of real numbers.

Let us consider two arbitrary real numbers a and b. Suppose that these numbers can be represented by the following infinite decimals:

$$a = \pm a_0, a_1 a_2 \ldots a_n \ldots, \tag{2.3}$$

$$b = \pm b_0, b_1 b_2 \ldots b_n \ldots, \tag{2.4}$$

where one sign out of the two $\pm$ is taken.

Two real numbers (2.3) and (2.4) are taken to be *equal* if they are of like signs and if the equalities $a_0 = b_0, a_1 = b_1, \ldots, a_n = b_n, \ldots$ are valid.

Suppose we have two *unequal* real numbers a and b. We shall now establish the rule which we can use to decide whether these numbers are connected by the $>$ sign or by the $<$ sign.

(1) First we shall suppose that the numbers a and b are *both nonnegative* and can be represented as $a = a_0, a_1 a_2 \ldots a_n \ldots$; $b = b_0, b_1 b_2 \ldots b_n \ldots$.

Since the numbers a and b are unequal, at least one of the equalities $a_0 = b_0, a_1 = b_1, \ldots, a_n = b_n, \ldots$ is violated. Let us denote by k the least number n for which the equality** $a_n = b_n$ is violated. Then we shall assume that $a > b$ if $a_k > b_k$ and $a < b$ if $a_k < b_k$.

(2) If *one* of the two numbers a and b is *nonnegative* and the *other* is *negative*, we shall naturally consider the nonnegative number to be larger than the negative number.

(3) It remains to consider the case when *both numbers a and b are negative. Let us agree to use the term the* **absolute value** *of a real number a for a nonnegative real number symbolized as* $|a|$ *and equal to the decimal fraction, representing the number a, taken with the plus sign.*

If a and b are *both negative*, we shall assume that $a > b$ if $|b| > |a|$, and $a < b$ if $|a| > |b|$.***

* This is quite in agreement with the process of measuring a line segment described in Sec. 2.1.2, for this process cannot lead to an infinite decimal fraction in which all decimal digits, beginning with a decimal place, are zeros.

** That is, we take it that $a_0 = b_0, a_1 = b_1, \ldots, a_{k-1} = b_{k-1}$, but $a_k \neq b_k$.

*** It is easy to see that the rule of comparison of real numbers formulated above, as applied to two rational numbers, leads to the same result as the rule of comparison of rational numbers given in the footnote on p. 36.

Indeed, it is sufficient to consider the case of two *nonnegative* rational numbers a and b. Let a be larger than b according to *the rule of comparison of rational*

Let us make sure that the rule of comparison of real numbers possesses property 1° formulated in 2.1.1 for rational numbers. Namely, we shall prove that if a, b and c are arbitrary real numbers and *if* $a > b$ *and* $b > c$, *then* $a > c$ (the transitivity of the sign $>$)*. To prove this property, we shall consider three possible cases.

(1) First suppose that $c \geqslant 0$. Then it is evident from the rule of comparison of real numbers that $b > 0$ and $a > 0$. Let $a = a_0, a_1a_2 \ldots a_n \ldots$; $b = b_0, b_1b_2 \ldots b_n \ldots$; $c = c_0, c_1c_2 \ldots c_n \ldots$. We shall denote by k the least number n for which the equality $a_n = b_n$ is violated (i.e. we shall assume that $a_0 = b_0$, $a_1 = b_1$, $\ldots$ $\ldots$, $a_{k-1} = b_{k-1}$, $a_k > b_k$), and by p we shall denote the least number n for which the equality $b_n = c_n$ is violated (i.e. we shall suppose that $b_0 = c_0$, $b_1 = c_1$, $\ldots$, $b_{p-1} = c_{p-1}$, $b_p > c_p$). Then, if we denote by m the least of the two numbers k and p, the relations $a_0 = c_0$, $a_1 = c_1$, $\ldots$, $a_{m-1} = c_{m-1}$, $a_m > c_m$ will hold true and this will mean than $a > c$.

(2) Let $c < 0$, $a \geqslant 0$. Then the relation $a > c$ will be valid for any b.

(3) It remains to consider the case when all the three numbers a, b and c are *negative*. Since $a > b$ and $b > c$, it follows that $|b| > |a|$ and $|c| > |b|$. But then, by virtue of the case of three positive numbers considered above, $|c| > |a|$, and this means that $a > c$. We have thus completely proved the transitivity of the $>$ sign.

2.1.4. Approximation of real numbers by rational numbers. In this subsection we shall show that any real number can be approximated by rational numbers with any degree of accuracy. Let us consider an arbitrary real number a. For definiteness, we shall consider this number to be nonnegative and shall represent it by an infinite decimal fraction $a = a_0, a_1a_2 \ldots a_n \ldots$

Terminating the indicated fraction at the n-th decimal place, we shall obtain a rational number $a_0, a_1a_2 \ldots a_n$. Increasing this number by $\frac{1}{10^n}$, we shall get another rational number $a_0, a_1a_2 \ldots$ $\ldots a_n + \frac{1}{10^n}$. It can easily be established, by the rule of comparison

numbers, and let $a = a_0, a_1a_2 \ldots a_n \ldots$; $b = b_0, b_1b_2 \ldots b_n \ldots$. Suppose a point M_1 on the number axis corresponds to the rational number a and a point M_2 to the rational number b. Then it is clear that point M_1 lies to the right of point M_2. At the same time it follows from 2.1.2 that the *integer* $a_0a_1 \ldots a_k$ $(b_0b_1 \ldots b_k)$ shows how many times $\frac{1}{10^k}$ of the unit length OE can be laid off along the line segment OM_1 (OM_2) with the right end deleted. Since the segment OM_1 is larger than the segment OM_2, there can be chosen a number k such that $a_0a_1 \ldots a_{k-1} = b_0b_1 \ldots b_{k-1}$, and $a_0a_1 \ldots a_k > b_0b_1 \ldots b_k$, but this precisely means that $a > b$ *according to the rule of comparison of real numbers.*

* The transitivity of the equality sign stating that if $a = b$ and $b = c$, then $a = c$ follows directly from the rule of comparison of real numbers.

of real numbers, that for any number n there hold inequalities

$$a_0,\ a_1a_2 \ldots a_n \leqslant a \leqslant a_0,\ \ a_1a_2 \ldots a_n + \frac{1}{10^n}. \tag{2.5}$$

Inequalities (2.5) imply that *the real number a is contained between two rational numbers whose difference is $\frac{1}{10^n}$. The number n can be arbitrary.*

Now we will show that for any *preassigned* positive rational number ε, beginning with a certain number n, there holds an inequality $\frac{1}{10^n} < \varepsilon$. Indeed, whatever the rational number $\varepsilon > 0$, there will be only a finite quantity of natural numbers not exceeding the number $\frac{1}{\varepsilon}$. Therefore, the inequality $10^n \leqslant \frac{1}{\varepsilon}$ or $\frac{1}{10^n} \geqslant \varepsilon$ is valid only for a finite quantity of the numbers n. For all the other numbers n there holds an inverse inequality $\frac{1}{10^n} < \varepsilon$, and that is what we intended to prove.

Thus we have come to the following *assertion: for any real number a and for any preassigned positive rational number ε there can always be chosen two rational numbers α_1 and α_2 such that $\alpha_1 \leqslant a \leqslant \alpha_2$, with $\alpha_2 - \alpha_1 < \varepsilon$.*

Inequalities (2.5) enable us to assert that the rational number $a_0, a_1a_2a_3 \ldots a_n$ approximates the real number a with an accuracy to within $\frac{1}{10^n}$. In practical applications we usually deal with approximate values of real numbers, replacing them by rational numbers with the required degree of accuracy.

2.1.5. Sets of real numbers bounded above and below. Here we shall consider an arbitrary set of real numbers containing at least one number.* It will be denoted as $\{x\}$. Separate numbers contained in the set $\{x\}$ will be called the *members* (*elements*) of that set.**

Definition 1. *A set of real numbers $\{x\}$ is said to be* **bounded above (below)** *if there exists a real number M (a number m) such that each element x of the set $\{x\}$ satisfies the inequality*

$$x \leqslant M \quad (x \geqslant m).$$

In this case, the number M (m) is called the *upper bound* (*lower bound*) of the set $\{x\}$.

Every set $\{x\}$ bounded above is sure to have an infinitude of upper bounds. Indeed, if the real number M is the upper bound of the set

---

* Such a set is usually called *nonempty*.

** The concept of a set and that of its element are basic concepts (see footnote on page 18).

$\{x\}$, then any real number M^* larger than M is also an upper bound of the set $\{x\}$. The argument concerning the lower bounds of the set $\{x\}$ bounded below is similar.

Thus, for instance, the set of all negative real numbers is bounded above. We can take any nonnegative real number as an upper bound M of such a set. The set of all integral positive numbers 1, 2, 3, . . . is bounded below. Any real number m satisfying the inequality $m \leqslant 1$ can serve as a lower bound of this set.

The question naturally arises as to the existence of the *least* upper bound of the set bounded above and the *greatest* lower bound of the set bounded below.

Definition 2. *The least of all the upper bounds of the set $\{x\}$ bounded above is called the* **supremum** *of that set and is symbolized as* $\bar{x} = \sup \{x\}$.

The greatest of all the lower bounds of the set $\{x\}$ bounded below is called the **infimum** *of that set and is symbolized as* $\underline{x} = \inf \{x\}$.

The second definition can be formulated differently, namely:

The number $\bar{x}$ (number $\underline{x}$) is the least upper (greatest lower) bound of the set $\{x\}$ bounded above (below), if the following two requirements are satisfied: (1) *each element x of the set $\{x\}$ complies with the inequality $x \leqslant \bar{x}$ ($x \geqslant \underline{x}$),* (2) *whatever the real number x' smaller than $\bar{x}$ (larger than $\underline{x}$), there can be found at least one element x of the set $\{x\}$ complying with the inequality $x > x'$ ($x < x'$).*

Requirement (1) implies that the number $\bar{x}$ (the number $\underline{x}$) is *one* of the upper (lower) bounds, and requirement (2) means that this bound is the *least* (*greatest*) and cannot be decreased (increased).

Evidently, a set of all negative real numbers has the least upper bound, the number zero, *this number not belonging to the indicated set.* The set of all integral positive numbers 1, 2, 3, . . . evidently has the greatest lower bound $\underline{x} = 1$, which *belongs to the indicated set* (i.e. is the *least element* of that set). Thus the least upper (greatest lower) bound of a set may belong and may not belong to that set.

The existence of the least upper (greatest lower) bound of any set bounded above (below) is not obvious and has to be proved.

Let us prove the following *fundamental* theorem.

Theorem 2.1. *If a set of real numbers contains at least one element and is bounded above (below), there exists a real number $\bar{x}$ (number $\underline{x}$) which is the least upper (greatest lower) bound of that set.*

Proof. We shall only prove the existence of the least upper bound of any set bounded above since the proof of the existence of the greatest lower bound of any set bounded below is quite analogous.

Thus, suppose the set $\{x\}$ is bounded above, which means that there exists a real number M such that each member x of the set $\{x\}$ satisfies

the inequality

$$x \leqslant M. \tag{2.6}$$

Here two cases are possible: 1°. Among the members of the set $\{x\}$ there is *at least one* nonnegative real number. 2°. All the members of that set are negative real numbers. We shall consider these two cases separately.

1°. We shall take only nonnegative real numbers belonging to the set $\{x\}$. Each of these numbers will be presented in the form of an infinite decimal fraction and the integral parts of these fractions will be considered. By virtue of (2.6) all the integral parts do not exceed the number M and, therefore, we shall take the *greatest* of these parts and denote it by $\bar{x}_0$. Of the nonnegative numbers of the set $\{x\}$ we shall retain those whose integral part is equal to $\bar{x}_0$ and shall delete all the others. In the retained numbers we shall consider the first decimal digits. We shall designate the greatest of these digits by $\bar{x}_1$. Of the nonnegative numbers of the set $\{x\}$ we shall retain those whose integral part is equal to $\bar{x}_0$ and the first decimal digit is equal to $\bar{x}_1$, and shall delete all the other numbers. In the retained numbers we shall consider the second decimal digits, designating the greatest of them by $\bar{x}_2$. Proceeding with analogous arguments, we shall successively determine the decimal digits of a certain real number $\bar{x}$:

$$\bar{x} = \bar{x}_0, \ \bar{x}_1\bar{x}_2 \ldots \bar{x}_n \ldots .$$

Let us prove that the real number $\bar{x}$ is precisely the least upper bound (supremum) of the set $\{x\}$. For our purpose it is sufficient to prove *two assertions*: (1) each member x of the set $\{x\}$ satisfies the inequality $x \leqslant \bar{x}$, (2) whatever the real number x', smaller than $\bar{x}$, there can be chosen at least one member x of the set $\{x\}$ satisfying the inequality $x > x'$.

Let us first prove *assertion* (1). Since $\bar{x}$ is nonnegative, any *negative* number x of the set $\{x\}$ is sure to satisfy the inequality $x \leqslant \bar{x}$. Let $x = x_0, x_1 \ldots x_n \ldots$ be any *nonnegative* number belonging to the set $\{x\}$.

Let us assume that this number x does not comply with the inequality $x \leqslant \bar{x}$. Then $x > \bar{x}$ and, in accordance with the rule of comparison, there is a number k such that $x_1 = \bar{x}_1, \ldots, x_{k-1} = \bar{x}_{k-1}, x_k > \bar{x}_k$. But the last relations are in contradiction with the fact that we take as $\bar{x}_k$ the *greatest* of the decimal digits x_k of the members x whose integral part and the first $(k-1)$ decimal digits are equal, respectively, to $\bar{x}_0, \bar{x}_1, \ldots, \bar{x}_{k-1}$.

Now we shall prove *assertion* (2). Let $x' = x'_0, x'_1 x'_2 \ldots x'_n \ldots$ be an arbitrary real number* smaller than $\bar{x}$. Then, in accordance with the rule of comparison of real numbers, there will be a number n such that

$$x'_0 = \bar{x}_0,\ x'_1 = \bar{x}_1,\ \ldots,\ x'_{n-1} = \bar{x}_{n-1},\ x'_n < \bar{x}_n. \tag{2.7}$$

On the other hand, we have taken the number $\bar{x}$ such that among the members of the set $\{x\}$ there is the number $x = x_0, x_1 x_2 \ldots x_n \ldots$, whose integral part and the first n decimal digits are the same as those of the number $\bar{x}$, that is

$$\bar{x}_0 = x_0,\ \bar{x}_1 = x_1,\ \ldots,\ \bar{x}_{n-1} = x_{n-1},\ \bar{x}_n = x_n. \tag{2.8}$$

Comparing (2.7) and (2.8), we shall obtain, by virtue of the rule of comparison of real numbers, $x' < x$. *We have proved assertion* (2). Thus, the existence of the supremum for case 1° has been proved.

2°. The proof of the existence of the supremum in the second case, when *all the members of the set* $\{x\}$ *are negative real numbers*, is analogous. In this case we shall represent all the members of the set $\{x\}$ as negative infinite decimals. Next, we shall denote by $\bar{x}_0$ *the least* of the integral parts of these fractions, by $\bar{x}_1$ *the least* of the first decimal digits of those fractions whose integral part is equal to $\bar{x}_0$, by $\bar{x}_2$, *the least* of the second decimal digits of those fractions whose integral part is equal to $\bar{x}_0$ and the first decimal digit is equal to $\bar{x}_1$, ... In this way we shall determine the negative real number

$$\bar{x} = -\bar{x}_0,\ \bar{x}_1 \bar{x}_2 \ldots \bar{x}_n \ldots .$$

By a complete analogy with case 1° it can be proved that the number $\bar{x}$ is the least upper bound of the set $\{x\}$, i.e. it satisfies two assertions formulated in the discussion of case 1°. Thus we have proved the theorem.

Remark. In proving theorem 2.1 for case 2°, all the decimal digits of the number $\bar{x} = -\bar{x}_0, \bar{x}_1 \bar{x}_2 \ldots \bar{x}_n \ldots$, beginning with some decimal place, may turn out to be equal to zero, i.e. this number may have the form $\bar{x} = -\bar{x}_0, \bar{x}_1 \ldots \bar{x}_{k-1} \bar{x}_k\, 000 \ldots$, where $\bar{x}_k \neq 0$. In such a case, the proof given above remains valid, but in accordance with the agreement accepted in 2.1.3, the number $\bar{x}$, when compared with the members of the set, must be written as

$$\bar{x} = -\bar{x}_0, \bar{x}_1 \ldots \bar{x}_{k-1} (\bar{x}_k - 1)\, 999 \ldots .$$

---

* Without loss of generality, we shall consider this number x' to be nonnegative, for were it negative, the inequality $x > x'$ would be satisfied by the nonnegative member x of the set $\{x\}$.

2.2. ARITHMETIC OPERATIONS ON REAL NUMBERS, BASIC PROPERTIES OF REAL NUMBERS

2.2.1. Definition of the sum of real numbers. One of the most significant problems of the theory of real numbers is *the defining of the operations of addition and multiplication of those numbers and the basic properties of these operations.* First, let us discuss the operation of addition of real numbers.

The technique of adding two real numbers is well known. Two real numbers a and b to be added are usually replaced by rational numbers with the required degree of accuracy, and the sum of these rational numbers is taken as an approximate value of the sum of the two given real numbers. In this case it is quite irrelevant whether the approximation is excessive or deficient. In fact, the indicated practical method of summation of real numbers implies that the more accurate the approximation (major or minor) of the rational numbers α and β to the real numbers a and b, respectively, the more accurate the approximation of the sum $\alpha + \beta$ to the real number which should be the sum of the real numbers a and b.

To substantiate the indicated technique of summation of real numbers, we shall give the following definition of the sum of two real numbers.

Let α_1 and α_2 be two arbitrary rational numbers containing between them a real number a (i.e. $\alpha_1 \leqslant a \leqslant \alpha_2$) and let β_1 and β_2 be arbitrary rational numbers containing between them a real number b (i.e. $\beta_1 \leqslant b \leqslant \beta_2$). Then we shall say that the sum of the real numbers a and b is a real number x which is contained between all the rational numbers $(\alpha_1 + \beta_1)$ and $(\alpha_2 + \beta_2)$.*

To put it otherwise, *the* **sum** *of the real numbers a and b is such a real number x which, for any rational numbers* α_1, α_2, β_1 *and* β_2 *satisfying the inequalities*

$$\alpha_1 \leqslant a \leqslant \alpha_2, \quad \beta_1 \leqslant b \leqslant \beta_2 \tag{2.9}$$

satisfies the inequalities

$$\alpha_1 + \beta_1 \leqslant x \leqslant \alpha_2 + \beta_2. \tag{2.10}$$

There is not the slightest doubt that such a real number x exists and that it is unique. The proof of this assertion is given below in small type. It is easy to ascertain that this number x is the least upper bound of the set $\{\alpha_1 + \beta_1\}$ of the sums of all rational numbers α_1 and β_1, satisfying the left-hand inequalities (2.9).**

* Note that in the elementary course the sum of two real numbers is usually defined in a similar way.

** We can prove by analogy that this number is the greatest lower bound of the set $\{\alpha_2 + \beta_2\}$ of the sums of all the rational numbers α_2 and β_2 satisfying the right-hand inequalities (2.9).

1°. First of all we shall make sure that the indicated upper bound does exist. Indeed, let us fix the arbitrary rational numbers α_2 and β_2 complying with the right-hand inequalities (2.9) and consider various rational numbers α_1 and β_1 satisfying the left-hand inequalities (2.9). From the transitivity of the sign $>$ established in 2.1.3 we conclude that $\alpha_1 \leqslant \alpha_2$, $\beta_1 \leqslant \beta_2$, and these inequalities yield an inequality $\alpha_1 + \beta_1 \leqslant \alpha_2 + \beta_2$ (see the end of 2.1.1). Thus, the set of all rational numbers $\{\alpha_1 + \beta_1\}$ is *bounded above* and the number $\alpha_2 + \beta_2$ is *one of the upper bounds* of that set. According to Theorem 2.1 the set $\{\alpha_1 + \beta_1\}$ has a supremum, which will be denoted by x. It remains to make sure that the number x is the sum of the real numbers a and b, that is it satisfies inequalities (2.10). In fact, according to the definition of the least upper bound, the left-hand inequality (2.10) is valid, and the validity of the right-hand inequality (2.10) follows from the fact that $\alpha_2 + \beta_2$ is *one of the upper bounds*, and x is the *least upper bound* of the set $\{\alpha_1 + \beta_1\}$.

2°. Now we shall ascertain that there is *only one* real number x satisfying inequalities (2.10). We shall make use of the following lemma (we think it more convenient to prove the lemma at the end of the present subsection).

If for two given real numbers x_1 and x_2 and for any preassigned positive rational ε there are two rational numbers γ_1 and γ_2 such that $\gamma_1 \leqslant x_1 \leqslant \gamma_2$, $\gamma_1 \leqslant x_2 \leqslant \gamma_2$ and $\gamma_2 - \gamma_1 < \varepsilon$, then the numbers x_1 and x_2 are equal.

Suppose there exist two real numbers x_1 and x_2 satisfying inequalities (2.10) (for any rational numbers α_1, α_2, β_1 and β_2 satisfying inequalities (2.9)). Let us take any positive rational number ε. According to the assertion proved in Sec. 2.1.4, there can be chosen for the real number a and for the rational number $\varepsilon/2$, rational numbers α_1 and α_2, such that $\alpha_1 \leqslant a \leqslant \alpha_2$, with $\alpha_2 - \alpha_1 < \varepsilon/2$. Similarly, for the real number b and for the rational number $\varepsilon/2$, there can be chosen rational numbers β_1 and β_2, such that $\beta_1 \leqslant b \leqslant \beta_2$, with $\beta_2 - \beta_1 < \varepsilon/2$.

In this way, both real numbers x_1 and x_2 will be contained between two rational numbers $(\alpha_1 + \beta_1)$ and $(\alpha_2 + \beta_2)$ whose difference in absolute values is equal to

$$(\alpha_2 + \beta_2) - (\alpha_1 + \beta_1) = (\alpha_2 - \alpha_1) + (\beta_2 - \beta_1) < \varepsilon.$$

Since ε is any preassigned positive rational number, it follows that $x_1 = x_2$, by virtue of the lemma formulated above.

3°. Let us now ascertain that, as applied to two *rational* numbers, the definition of the sum of real numbers formulated above and the definition of the sum of rational numbers known from the elementary course *lead to the same result.* Indeed, if a and b are two rational numbers satisfying the inequalities $\alpha_1 \leqslant a \leqslant \alpha_2$, $\beta_1 \leqslant b \leqslant \beta_2$ and $(a + b)$ is their sum obtained according to the definition familiar from the elementary course, then, evidently,

$$(\alpha_1 + \beta_1) \leqslant a + b \leqslant (\alpha_2 + \beta_2) \tag{2.11}$$

and, according to the assertion just proved, the rational number $(a + b)$ is the only real number satisfying inequalities (2.11).

4°. Before proving the lemma formulated above, we shall make the following auxiliary assertion.

Whatever the two real numbers a and b such that $b < a$, there can be chosen a rational number α contained between them, that is a number satisfying the inequality $b < \alpha < a$ (and, consequently, there can be found an infinitude of various rational numbers contained between a and b).

It is, evidently, sufficient to consider the case when both numbers a and b are *nonnegative*, for the case of both numbers a and b being *nonpositive* can be reduced to the indicated case by passing to the absolute values and the case when *one number is positive* and *the other negative* is trivial (we can take zero as the number α). And so we postulate that $b \geqslant 0$; $b < a$; $a = a_0, a_1 a_2 \ldots a \ldots$; $b = b_0, b_1 b_2 \ldots b_n \ldots$. Let k be the least of the numbers n for which the equality $a_n = b_n$ is violated, i.e. $a_0 = b_0$, $a_1 = b_1$, ..., $a_{k-1} = b_{k-1}$, $a_k > b_k$.

As we have agreed in 2.1.3, *all the a_n cannot be equal to zero when $n > k$.* Let p be *the least* of the numbers n exceeding k for which $a_n > 0$, i.e.

$$a = a_0, a_1 \ldots a_k 00 \ldots 0 a_p \ldots$$

Then it follows directly from the rule of comparison of real numbers that the rational number $\alpha = a_0, a_1 \ldots a_k 00 \ldots 0 (a_p - 1) 999 \ldots$ satisfies the inequalities $b < \alpha < a$. We have proved the auxiliary assertion.

Turning now to the proof of the lemma, we shall assume that $x_1 \neq x_2$. For definiteness, we shall suppose that $x_1 < x_2$. Then in accordance with the auxiliary assertion, there will exist two rational numbers α_1 and α_2 such that

$$x_1 < \alpha_1 < \alpha_2 < x_2. \tag{2.12}$$

Now let γ_1 and γ_2 be two arbitrary rational numbers satisfying the inequalities

$$\gamma_1 \leqslant x_1 \leqslant \gamma_2, \quad \gamma_1 \leqslant x_2 \leqslant \gamma_2. \tag{2.13}$$

From the comparison of (2.12) and (2.13) and from the transitivity of the sign $>$ it follows that $\gamma_1 < \alpha_1 < \alpha_2 < \gamma_2$. But then $\gamma_2 - \gamma_1 > \alpha_2 - \alpha_1$ which contradicts the fact that the difference $\gamma_2 - \gamma_1$ can be made smaller than any preassigned positive rational number ε. Thus we have proved the lemma.

2.2.2. Definition of the product of real numbers. Since the problems which arise when we define a product of real numbers mainly coincide with those considered in the process of defining a sum of real numbers, we shall be brief here, giving only the results.

We shall first define the product of two *positive* numbers a and b. For that purpose we shall designate by α_1, α_2, β_1 and β_2 any positive rational numbers satisfying the inequalities $\alpha_1 \leqslant a \leqslant \alpha_2$, $\beta_1 \leqslant b \leqslant \beta_2$.

The product of the *positive* real numbers a and b will be the real number x complying with the inequalities $\alpha_1\beta_1 \leqslant x \leqslant \alpha_2\beta_2$.

Just as in the case of the sum, we establish that such a real number x does exist and is unique. It is easy to ascertain that such a number x is the supremum of the set $\{\alpha_1\beta_1\}$ of the products of all rational numbers α_1 and β_1 satisfying the inequalities $0 < \alpha_1 \leqslant a$, $0 < \beta_1 \leqslant b$.

The product of real numbers of *any sign* can be determined according to the following rule:

(1) it is assumed that $a \cdot 0 = 0 \cdot a = 0$;

(2) it is assumed that $ab = \begin{cases} |a| \cdot |b|, & \text{if } a \text{ and } b \text{ are of like signs,} \\ -|a| \cdot |b|, & \text{if } a \text{ and } b \text{ are of unlike signs.} \end{cases}$

In conclusion we want to point out that just as in the case of a sum, it can be proved that, as applied to two rational numbers, the definition of the product of real numbers and the definition of the product of rational numbers known from the elementary course lead to the same results.

2.2.3. Properties of real numbers. We shall make sure now that the *basic properties* of rational numbers enumerated in 2.1.1 are also valid for arbitrary real numbers. The validity of property 1° for

real numbers has been established above. So it remains to establish the validity for real numbers of properties 2°-13°.

It is easy to verify the validity for real numbers of properties 2°-5° and 11° since they are related to the concept of the sum. The validity of properties 2°-5° follows directly from the definition of the sum of real numbers and from the validity of the indicated properties for rational numbers.

We shall now consider property 11°, that is, we shall prove that *if a, b and c are any three real numbers and $a > b$, then $a + c > b + c$.*

Since $a > b$, the auxiliary assertion made in the process of proving the lemma (at the end of 2.2.1) enables us to state that there can be chosen rational numbers α_1 and β_2 such that $a > \alpha_1 > \beta_2 > b$. For the real number c and for the positive rational number $\varepsilon = \alpha_1 - \beta_2$ there can be chosen rational numbers γ_1 and γ_2 such that $\gamma_1 \leqslant c \leqslant \gamma_2$, with $\gamma_2 - \gamma_1 < \varepsilon = \alpha_1 - \beta_2$ (see the assertion proved in 2.1.4).

Let then α_2 and β_1 be any rational numbers satisfying the inequalities $\alpha_2 \geqslant a$, $b \geqslant \beta_1$. Then according to the definition of the sum of real numbers, we have

$$\alpha_2 + \gamma_2 \geqslant a + c \geqslant \alpha_1 + \gamma_1, \quad \beta_2 + \gamma_2 \geqslant b + c \geqslant \beta_1 + \gamma_1.$$

To prove that $a + c > b + c$, it is sufficient to prove, by virtue of the transitivity of the sign $>$, that $\alpha_1 + \gamma_1 > \beta_2 + \gamma_2$, but this follows directly from the inequality $\gamma_2 - \gamma_1 < \alpha_1 - \beta_2$.

On the basis of properties 2°-5°, we can assert with complete assurance that *subtraction* of real numbers is an operation inverse of summation. *We shall say that the* **difference** *of the real numbers a and b is a real number c such that $c + b = a$.*

We shall make sure that this difference is the number $c = a + b'$, where b' is the number inverse of b.

Indeed, using properties 2°-5°, we can write

$$c + b = (a + b') + b = a + (b' + b) = a + 0 = a.$$

Now we shall ascertain that there is *only one* real number which is the difference between two given real numbers. Suppose that besides the number $c = a + b'$ indicated above, there exists one more number d such that $d + b = a$. Then, on one hand, $(d + b) + b' = a + b' = c$, and, on the other hand, $(d + b) + b' = d + (b + b') = d + 0 = d$, i.e. $c = d$.

It follows from the definition of the difference and from property 5° that the number a', inverse of a, is equal to the difference between the numbers 0 and a. This number is usually written as $-a$.

Properties 6°, 7°, 8°, 9°, 10° and 12° related to the concept of the product can be easily extended to the case of real numbers. It should be only noted, in connection with property 9°, that if a is a positive real number and α_1 and α_2 are arbitrary rational numbers complying with the inequalities $0 < \alpha_1 \leqslant a \leqslant \alpha_2$, then the number a', inverse

of a, is said to be the only real number satisfying the inequalities $\frac{1}{\alpha_2} \leqslant a' \leqslant \frac{1}{\alpha_1}$.*

Properties 6°-9° enable us to infer that for any two real numbers a and b ($b \neq 0$) there exists a real number c satisfying the condition $cb = a$, and that this number is unique. That number c is referred to as the quotient of the numbers a and b.

The definition of the quotient and property 9° imply that the number a', inverse of a, is equal to the quotient of the numbers 1 and a which we will denote as $\frac{1}{a}$.

Let us finally note that property 13° of rational numbers can be reformulated for the case of real numbers as follows: *whatever the real number a, the number 1 can be taken as a summand as many times as is necessary to obtain the sum exceeding a.*** Let us prove this property. When $a < 0$, the proof is unnecessary since $1 > a$. Assume that $a \geqslant 0$; $a = a_0, a_1 \ldots a_n \ldots$. Since the definition of the sum of real numbers, in the particular case when these numbers are rational, coincides with the definition of the sum of rational numbers, we shall repeat the number 1 as a summand n times and obtain an integer n. Thus, it is sufficient to prove that for the number a there is an integer n such that $n > a$. But that is evident: it suffices to take $n = a_0 + 2$.

Thus we can say that *all the basic properties formulated for rational numbers in 2.2.1 can be applied to the case of real numbers. Consequently, all the rules of algebra related to the arithmetical operations and to the combination of equalities and inequalities remain valid for the case of real numbers.*

Here we shall end our exposition of the elements of the theory of real numbers which are necessary to construct the course of mathematical analysis. Further development of the theory of real numbers can be found in the Appendix, at the end of the book.

We must note in conclusion that we have constructed the theory of real numbers making use of their representation as infinite decimal fractions.

It is quite evident that we could make use of infinite fractions with any other, not necessarily decimal, base. In this respect, the number systems with different bases are equivalent. However, in certain problems connected with approximate calculations, and, in particular, in the cases where the numbers are rounded off to the specified digit, the number systems with even and odd bases behave quite

* We can take as the numder a' the supremum of the set of all rational numbers $\left\{\frac{1}{\alpha_2}\right\}$.

** This property is referred to as the Archimedes axiom.

differently (in this connection see Supplement 2 to the present chapter).

2.2.4. Some frequently employed relations. Let us prove the validity of the following two relations for any real numbers a and b:

$$|ab| = |a| \cdot |b|, \tag{2.14}$$

$$|a + b| \leqslant |a| + |b|. \tag{2.15}$$

These formulas can be put into words as follows: (1) *the absolute value of the product of two numbers is equal to the product of the absolute values of these numbers*, (2) *the absolute value of the sum of two numbers does not exceed the sum of the absolute values of these numbers.*

Relation (2.14) follows directly from the definition of the product of two real numbers. Let us prove relation (2.15). Proceeding from the definition of the absolute value and the rule of comparison, we can state that the inequalities

$$-|a| \leqslant a \leqslant |a|, \quad -|b| \leqslant b \leqslant |b|$$

are valid for any real numbers a and b. By virtue of the basic properties, inequalities of like signs can be added termwise (this was proved at the end of 2.1.1). Therefore

$$-(|a| + |b|) \leqslant a + b \leqslant |a| + |b|.$$

Making use of the right-hand inequality in the case $a + b > 0$ and the left-hand inequality in the case $a + b \leqslant 0$, we obtain inequality (2.15).

Remark. Note another two frequently used inequalities:

$$|a - b| \geqslant |a| - |b| \tag{2.16}$$

and

$$|a - b| \geqslant ||a| - |b||. \tag{2.17}$$

To obtain inequality (2.16) it is sufficient to take into account that $a = (a - b) + b$ and, proceeding from (2.15), to write the inequality $|a| \leqslant |a - b| + |b|$. Inequality (2.17) follows from inequality (2.16) and from the inequality $|b - a| \geqslant |b| - |a|$, which results from (2.16) when a and b exchange their roles.

2.3. SOME CONCRETE SETS OF REAL NUMBERS

In the various sets of real numbers dealt with in what follows we shall designate as $\{x\}$ arbitrary sets of real numbers and shall refer to the numbers contained in them as the *members* or *points* of these sets. We shall say that *point* x_1 *of the set* $\{x\}$ *differs from point* x_2 *of that set if the real numbers* x_1 *and* x_2 *are not equal to each other.* If, in that case, the inequality $x_1 > x_2$ ($x_1 < x_2$) is valid, we shall say that *point* x_1 *lies to the right* (*to the left*) *of point* x_2.

Now we shall consider some of the most frequently used sets of real numbers.

1°. The set of real numbers x satisfying the inequalities $a \leqslant x \leqslant b$, where $a < b$, will be referred to as a *closed interval* and symbolized as $[a, b]$. In that case, the numbers a and b will be called the *end points* of the interval $[a, b]$, and an arbitrary number x satisfying the inequalities $a < x < b$, the *interior point* of the interval $[a, b]$.

2°. The set of all real numbers x satisfying the inequalities $a \leqslant x < b$ (or $a < x \leqslant b$) will be referred to as a *half-open (half-closed) interval* and symbolized as $[a, b)$ {or $(a, b]$}.

3°. The set of all real numbers x satisfying the inequalities $a < x < b$ will be referred to as an *open interval* (or simply *interval*) and symbolized as (a, b).

4°. Any interval containing a point c will be called the *neighbourhood of point c.*

5°. The interval $(c - \varepsilon, c + \varepsilon)$, where $\varepsilon > 0$, will be called the *ε-neighbourhood of point c.*

6°. The set of all real numbers will be called an *infinite interval* or an *(infinite) number line* and symbolized as $(-\infty, +\infty)$.

7°. The set of all real numbers x satisfying the inequality $x \geqslant a$ (or $x \leqslant b$) will be called a *half-line* and symbolized as $[a, \infty)$ {or $(-\infty, b]$}.

8°. The set of all real numbers x, satisfying the inequality $x > a$ (or $x < b$) will be called an *open half-line* and symbolized as (a, ∞) {or $(-\infty, b)$}.

Remark. An arbitrary set $\{x\}$ will be referred to as *dense in itself*, if any neighbourhood of each point x of that set contains at least one point of the set different from x. Any of the sets defined in 1°-8° can serve as an example of a set dense in itself. Another example of a set dense in itself may be any set of all rational numbers included in any of the sets 1°-8°.

SUPPLEMENT 1 TO CHAPTER 2

DECIMAL-TO-BINARY CONVERSION AND BINARY-TO-DECIMAL CONVERSION OF NUMBERS

In this supplement we shall consider the algorithms of decimal-to-binary conversion of numbers and the converse process.

1. Decimal-to-binary conversion of numbers. To specify the decimal number x_{10} in the location of a computer, use is made of the so-called *normalized* notation of that number:

$$x_{10} = q_{10} \cdot 10^{p_{10}}. \tag{2.18}$$

In this notation the quantity

$$q_{10} = (1 - 2S_q)(\alpha_1 \cdot 10^{-1} + \alpha_2 \cdot 10^{-2} + \ldots + \alpha_9 \cdot 10^{-9}) \tag{2.19}$$

is called the decimal *mantissa* of the given number, with $\alpha_1 \geqslant 1$, S_q being taken equal to zero for $q_{10} \geqslant 0$ and equal to unity for $q_{10} < 0$,* and the *exponent*

$$p_{10} = (1 - 2S_p)(\beta_1 + 10\beta_2) \tag{2.20}$$

is called the decimal *order* of the given number, with S_p being taken equal to zero if $p_{10} \geqslant 0$ and equal to unity if $p_{10} < 0$.**

Figure 2.3 shows how a decimal number (2.18)-(2.20) can be specified in the locations of a computer. There are four binary digits to represent each of the decimal numbers β_1, α_1, α_2, ..., α_9 so that each of the indicated numbers can

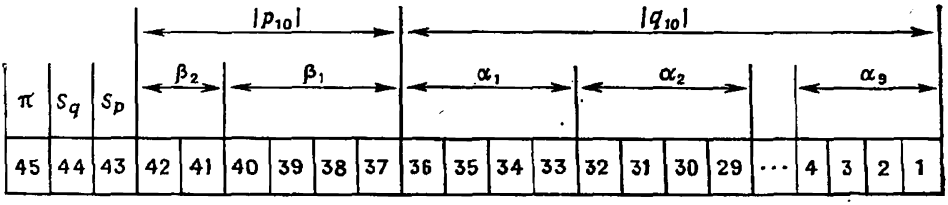


Fig. 2.3

take on any of the integral values from 0 to 15, while there are only two digits to represent the number β_2 so that β_2 can take on the values 0, 1, 2, 3.***

The standard program generates for the decimal number (2.18)-(2.20) a respective binary number x_2. This program is realized as follows. First the quantity γ is computed:

$$\gamma = (1 - 2S_q)(\alpha_1 \cdot 10^9 + \alpha_2 \cdot 10^8 + \ldots + \alpha_9) \cdot 2^{68}.$$

Then the computed quantity γ is multiplied by $k = 2^{68} \cdot 10^{-10}$ (the latter quantity is specified in the computer in a normalized form as well, usually with an excess of two units of the least significant digit of the mantissa).

The product $k\gamma$ evidently corresponds to the decimal mantissa (2.19). The next procedure consists in the multiplication of $k\gamma$ by 10 or 1/10 depending on the sign of p_{10} (i.e. on S_p) performed as many times as is the value of $|p_{10}|$. The program is usually completed by clearing the three least significant digits of the mantissa.

The program discussed (1) ensures at least 30 correct binary digits in the result, (2) makes it possible to convert any integral decimal number from 0 to 50 000 into a binary number, (3) ensures a conversion of a normalized decimal number into a normalized binary number.****

---

* Thus, the factor $(1-2S_q)$ in equality (2.19) characterizes the sign of the mantissa q_{10}.

** So that the factor $(1 - 2S_p)$ in equality (2.20) characterizes the sign of the exponent p_{10}.

*** However, because of the construction of the keyboard, the highest value of each of the numbers β_1, α_1, α_2, ..., α that can be punched in each position cannot be more than nine, and that of the number β_2 cannot be more than unity. Thus, each of the numbers β_1, α_1, α_2, ..., α_9 varies in the range from 0 to 9 and the number β_2 takes on the values 0 and 1.

**** If the decimal number taken was not normalized (i.e. in (2.19) the condition $\alpha_1 \geqslant 1$ was violated), the result of its conversion into a binary notation may prove to be not normalized.

We should note in conclusion that if during the realization of the indicated program, the binary order of the number being converted exceeds 59 in the process of multiplication by 10, the multiplication by 10 is terminated even if it is required by the value of $| p_{10} |$.

2. Binary-to-decimal conversion of numbers .We wish to indicate a standard program according to which a given binary number $x_2 = q_2 \cdot 2^{p_2}$ written in a normalized form can be used to obtain a corresponding decimal number x_{10} written in the normalized form (2.18)-(2.20). The number is written in the locations of the computer as shown in Fig. 2.3.

The program is realized as follows. First the initial number x_2 is multiplied by 1/10 to obviate the word size overflow upon the following multiplication by 10. Then, the number obtained is multiplied by 10 or 1/10, depending on whether it is greater than unity or not, until the result of the multiplications gets into the range from 1/10 to 1. The number of multiplications is evidently defined by $| p_{10} |$. As to the sign of p_{10}, it is positive if the initial number is greater than unity and negative if otherwise.

It is evident that the number obtained as a result of the multiplications is the decimal mantissa q_{10}. The digits $\alpha_1, \alpha_2, \ldots, \alpha_9$ of the decimal mantissa q_{10} are obtained in a consecutive order, by multiplying by 10 and separating the integral part.

SUPPLEMENT 2 TO CHAPTER 2

ON THE ERRORS IN ROUNDING OFF THE NUMBERS IN THE NUMBER SYSTEM WITH EVEN AND ODD BASES

Suppose a computer works with t-digit numbers in the number system with the base $p \geqslant 2$. Then, without any loss of generality, we can assume that all the numbers $x^{(t)}$ stored in the computer memory have the form

$$x^{(t)} = a_1 p^{-1} + a_2 p^{-2} + \ldots + a_t p^{-t},$$

where the coefficients a_i ($i = 1, 2, \ldots, t$) may take on the values $0, 1, \ldots, (p - 1)$. It is quite clear that such operations as addition, multiplication or division carried out on t-digit numbers may result in numbers consisting of more than t digits, and so a necessity arises to round off these numbers to t digits.

Let us consider a simple operation of rounding off the numbers consisting of $t + r$ digits (where $r > 0$) to the numbers consisting of t digits. Whatever the method used to round off the number $x^{(t+r)}$ containing $(t + r)$ digits, the result of the rounding must be a t-digit number. Hence it follows that the error in rounding off the number $x^{(t+r)}$ (we shall symbolize this error as $\varepsilon (x^{(t+r)})$) has the following form:

$$\varepsilon (x^{(t+r)}) = - ip^{-(t+r)} + m(i) \cdot p^{-t}.$$

In this formula i can take on the values $0, 1, \ldots, p^r - 1$ depending on the values of the last r digits of the number $x^{(t+z)}$, and $m(i)$ is a certain function of i taking on integral values and dependent on the chosen method of rounding off.

The most important characteristic of the error in rounding off is its mean value Δ, which is specified as a fraction*

$$\Delta = \frac{\sum \varepsilon\,(x^{(t+r)})}{n^{(t+r)}},$$

whose numerator is the sum of the errors corresponding to all the admissible values of the number $x^{(t+r)}$ and the denominator expresses the quantity of these numbers $x^{(t+r)}$.

Suppose all the numbers $x^{(t+r)}$ considered satisfy the inequalities $0 \leqslant x^{(t+r)} < 1$. Then, evidently, the quantity $n^{(t+r)}$ of all the numbers $x^{(t+r)}$ will be equal to $p^{(t+r)}$, and after some simple calculations we shall have

$$\Delta = \frac{\sum [-ip^{-(t+r)} + m(i) \cdot p^{-t}]}{p^{t+r}} = p^{-(t+r)} \left\{ \sum_{i=0}^{p^r-1} m\,(i) - \frac{p^r-1}{2} \right\}.$$

The sum $\sum m\,(i)$ in the braces depends on the method of rounding off, but in any case this sum is integral-valued. The second member $\frac{p^r-1}{2}$ in the braces will not be integral for any *even* p. Thus, for any even base p the mean error Δ is not equal to zero. This means that for any fixed method of rounding off consisting in deleting digits, the error in rounding off to the less number of digits will have a systematic shift for any number system with an even base. On the other hand, it can be easily verified that ordinary rounding-off in any number system with an odd base, adopted at school, leads to "unshifted" errors.

---

* The symbol $\sum$ is the sign of summation of the summands following this symbol. If these summands depend on the number i, the notation $\sum_{i=m}^{n}$ means that the summation must be performed over all the values of i, from m to n.

CHAPTER 3

LIMIT OF A SEQUENCE

One of the main operations of mathematical analysis is a limiting process. There are various forms of this process encountered in the analysis. In the present chapter we shall consider a simple form of this operation based on the concept of the limit of the so-called number sequence. The concept of the limit of a number sequence will enable us, in our further discussion, to define other forms of the limiting process.

3.1. NUMBER SEQUENCES

3.1.1. Number sequences and operations on them. The reader must be familiar with the concept of number sequences from elementary mathematics. We may cite as examples of number sequences (1) the sequence of all terms of an arithmetic or geometric progression, (2) the sequence of the perimeters of regular n-gons inscribed in a given circle, (3) the sequence $x_1 = 1$, $x_2 = 1.4$, $x_3 = 1.41$ . . . of the approximate values of the number $\sqrt{2}$. We shall begin this subsection with the elucidation of the concept of a number sequence.

If, in accordance with a definite rule, every number n of a series of natural numbers 1, 2, . . ., n . . . *is put into correspondence with a certain real number* x_n, *then the set of the numbered real quantities*

$$x_1, x_2, \ldots, x_n, \ldots \tag{3.1}$$

is referred to as a number sequence or simply a sequence.

The numbers x_n will be referred to as *terms* of the sequence (3.1). In an abbreviated form sequence (3.1) may be symbolized as $\{x_n\}$. Thus, for instance, we shall use the symbol $\left\{\frac{1}{n}\right\}$ for the sequence 1, 1/2, . . ., $1/n$, . . . and the symbol $\{1+(-1)^n\}$ for the sequence 0, 2, 0, 2, ...

We shall now acquaint the reader with arithmetic operations performed on the number sequence. Suppose we are given arbitrary sequences $x_1, x_2, \ldots, x_n, \ldots$ and $y_1, y_2, \ldots, y_n, \ldots$. The sequence $x_1+y_1, x_2+y_2, \ldots, x_n+y_n, \ldots$ (or $\{x_n+y_n\}$) will be

referred to as the *sum* of these sequences, the sequence x_1-y_1, x_2-y_2, ..., x_n-y_n, ... (or $\{x_n-y_n\}$) as the *difference* between these sequences, the sequence $x_1 \cdot y_1$, $x_2 \cdot y_2$, ..., $x_n \cdot y_n$, ... (or $\{x_n \cdot y_n\}$) as their *product*, and the sequence $\frac{x_1}{y_1}$, $\frac{x_2}{y_2}$, ... $\frac{x_n}{y_n}$, ... $\left(\text{or } \left\{\frac{x_n}{y_n}\right\}\right)$ as their *quotient*.

Remark. When determining the quotient $\left\{\frac{x_n}{y_n}\right\}$ it is required that *all the terms* y_n of the sequence $\{y_n\}$ be different from zero. However, if *only a finite number of the terms* of the sequence $\{y_n\}$ vanishes, the quotient $\left\{\frac{x_n}{y_n}\right\}$ can be determined beginning with the number starting from which all the terms y_n are different from zero.

3.1.2. Bounded and unbounded sequences.

Definition 1. *A sequence $\{x_n\}$ is said to be* **bounded above (below),** *if there exists a real number M (m) such that every term x_n of the sequence $\{x_n\}$ satisfies the inequality $x_n \leqslant M$ ($x_n \geqslant m$).**

In this case the number M (number m) is called the *upper bound* (*lower bound*) of the sequence $\{x_n\}$, and the inequality $x_n \leqslant M$ ($x_n \geqslant m$) is referred to as the *condition of boundedness* of the sequence from above (from below).

We wish to attract attention to the fact that any sequence $\{x_n\}$ bounded above has an infinitude of upper bounds. Indeed, if M is an upper bound, then any number M^* larger than M is also an upper bound. We want to emphasize that in the condition $x_n \leqslant M$ of boundedness of the sequence $\{x_n\}$ from above any of the upper bounds may be taken as M. Similar remarks can be made concerning the lower bounds of the sequence $\{x_n\}$ bounded from below.

Definition 2. *The sequence $\{x_n\}$ is said to be* **bounded** *if it has both the upper bound and the lower bound, i.e. if there exist numbers m and M such that any term x_n of that sequence satisfies the inequalities $m \leqslant x_n \leqslant M$.*

If the sequence $\{x_n\}$ is bounded and M and m are its upper bound and lower bound respectively, all the terms x_n of that sequence satisfy the inequality

$$|x_n| \leqslant A, \tag{3.2}$$

where A is the maximal of the two numbers $|M|$ and $|m|$. Conversely, if all the terms of the sequence $\{x_n\}$ satisfy inequality (3.2), then the inequalities $-A \leqslant x_n \leqslant A$ are satisfied too, and, conse-

---

* This definition is completely analogous to the definition of the set of real numbers bounded above (below) (see 2.1.5).

quently, the sequence $\{x_n\}$ is bounded. Thus, inequality (3.2) is another form of the condition of boundedness of a sequence. Let us elucidate the concept of an unbounded sequence. *The sequence* $\{x_n\}$ *is said to be unbounded, if for any positive number* A *there can be found a term* x_n *of this sequence satisfying the inequality* $|x_n| > A$. Let us consider some examples.

(1) The sequence $-1, -4, -9, \ldots, -n^2, \ldots$ has an upper bound and has no lower bound. The upper bound of this sequence is any number not smaller than -1.

(2) The sequence $1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}, \ldots$ is bounded. Indeed, any number $M \geqslant 1$ can serve as an upper bound of this sequence, and any number $m \leqslant 0$ can serve as its lower bound.

(3) The sequence $1, 2, 1, 3, \ldots, 1, n, 1, (n+1), \ldots$ is unbounded. In fact, whatever the positive number A, there can be found, among the terms of that sequence (with even numbers), some terms exceeding A.

3.1.3. Infinitely large and infinitesimal sequences.

Definition 1. *The sequence* $\{x_n\}$ *is called* **infinitely large** *if for any positive number* A* *there can be found a number* N** *such that for* $n \geqslant N$ *all the terms* x_n *of that sequence satisfy the inequality* $|x_n| > A$.

Remark. It is evident that any infinitely large sequence is unbounded, since for any $A > 0$, there can be chosen a number N such that when $n \geqslant N$, *all the terms* x_n satisfy the inequality $|x_n| > A$, and, consequently, for any $A > 0$ there can be found at least one term x_n such that $|x_n| > A$. However, an unbounded sequence is not necessarily infinitely large. For instance, the unbounded sequence $1, 2, 1, 3, \ldots, 1, n, \ldots$ is not infinitely large, since for $A > 1$ the inequality $|x_n| > A$ does not hold *for all* x_n with odd indices.

Definition 2. *The sequence* $\{\alpha_n\}$*** *is said to be* **infinitesimal** *if for any positive number* ε**** *there can be found a number* N***** *such that when* $n \geqslant N$ *all the terms* α_n *of that sequence satisfy the inequality* $|\alpha_n| < \varepsilon$.

Let us consider the following examples.

(1) We shall prove that for $|q| > 1$ the sequence $q, q^2, q^3, \ldots, q^n, \ldots$ is infinitely large and for $|q| < 1$ it is infinitesimal.

We shall first discuss the case $|q| > 1$. In this case $|q| = 1 + \delta$, where $\delta > 0$. Making use of Newton's binomial formula, we obtain $|q|^N = (1+\delta)^N = 1 + \delta N +$ (positive terms). Hence

$$|q|^N > \delta N. \tag{3.3}$$

---

* However large we take it.

** Since the number N depends on A, it is sometimes written as $N = N(A)$.

*** The terms of infinitesimal sequences are usually denoted by Greek letters.

**** However small we take it.

***** Since the number N depends on ε, it is sometimes written as $N = N(\varepsilon)$.

Let us fix an arbitrary number $A > 0$ and choose a number N sufficiently large of the inequality $\delta N > A$* to hold. The last inequality and inequality (3.3) yield an inequality $|q|^N > A$. Since $|q|^n \geqslant |q|^N$ for $n \geqslant N$ and for $|q| > 1$ (by virtue of the properties of the product of real numbers), we have $|q|^n > A$ for $n \geqslant N$. Thus we have proved that at $|q| > 1$ the sequence under consideration is infinitely large.

The case $|q| < 1$ is quite analogous. Here $\frac{1}{|q|} = 1 + \delta$, where $\delta > 0$ (we have omitted the case when $q = 0$). Using again Newton's binomial formula, we shall obtain, instead of (3.3), the following inequality:

$$\frac{1}{|q|^N} > \delta N \quad \text{or} \quad |q|^N < \frac{1}{\delta N}. \tag{3.3*}$$

Having fixed an arbitrary $\varepsilon > 0$, we shall choose a number N under the condition that $\frac{1}{\delta N} < \varepsilon$**. Since $|q|^n \leqslant |q|^N$ for $n \geqslant N$ and for $|q| < 1$, it follows from the inequalities obtained that $|q|^n < \varepsilon$ for $n \geqslant N$. Thus we have proved that at $|q| < 1$ the sequence under consideration is infinitesimal.

(2) Let us prove that the sequence $1, \frac{1}{2}, \ldots, \frac{1}{n}, \ldots$ is infinitesimal. Indeed, if $n \geqslant N$, then $\frac{1}{n} \leqslant \frac{1}{N}$. Therefore, proceeding from the given ε, it is sufficient to choose a number N under the condition that $\frac{1}{N} < \varepsilon$. For instance, we can put $N = \left[\frac{1}{\varepsilon}\right] + 1$.

3.1.4. Basic properties of infinitesimal sequences.

Theorem 3.1. *The sum of two infinitesimal sequences is an infinitesimal sequence.*

Proof. Let $\{\alpha_n\}$ and $\{\beta_n\}$ be infinitesimal sequences. We shall prove that the sequence $\{\alpha_n + \beta_n\}$ is infinitesimal. Assume that ε is an arbitrary positive number, N_1 is a number beginning with which $|\alpha_n| < \frac{\varepsilon}{2}$, and N_2 is a number beginning with which $|\beta_n| < \frac{\varepsilon}{2}$. (Such numbers N_1 and N_2 can be found in accordance with the definition of an infinitesimal sequence.) Since the absolute value of the sum of two numbers does not exceed the sum of their absolute values, i.e. $|\alpha_n + \beta_n| \leqslant |\alpha_n| + |\beta_n|$ (see 2.2.4), denot-

---

* It it sufficient to put $N = \left[\frac{A}{\delta}\right] + 1$, where $[x]$ denotes the integral part of the number x. For instance, $[5.138] = 5$, $[-172.9] = -173$.

** It is sufficient to put $N = \left[\frac{1}{\delta\varepsilon}\right] + 1$.

ing by N the largest of the two numbers, N_1 and N_2, we shall see that beginning with the number N there holds an inequality $|\alpha_n + \beta_n| < \varepsilon$. This means that the sequence $\{\alpha_n + \beta_n\}$ is infinitesimal and the proof of the theorem is complete.

Theorem 3.2. *The difference between two infinitesimal sequences is an infinitesimal sequence.*

The proof of this theorem is quite analogous to that given above, with the only difference that instead of the inequality $|\alpha_n + \beta_n| \leqslant |\alpha_n| + |\beta_n|$ we must take the inequality $|\alpha_n - \beta_n| \leqslant |\alpha_n| + |\beta_n|$.

Corollary. *The algebraic sum of any finite number of infinitesimal sequences is an infinitesimal sequence.*

Theorem 3.3. *An infinitesimal sequence is bounded.*

Proof. Let $\{\alpha_n\}$ be an infinitesimal sequence and ε a certain positive number. Furthermore, let N be a number beginning with which $|\alpha_n| < \varepsilon$. We shall denote by A the largest of the following N numbers: $\varepsilon, |\alpha_1|, |\alpha_2|, \ldots, |\alpha_{N-1}|$. We can write it as $A = \max\{\varepsilon, |\alpha_1|, |\alpha_2|, \ldots, |\alpha_{N-1}|\}$.* Evidently, $|\alpha_n| \leqslant A$ *for any number* n, and this implies the boundedness of the sequence, which proves the theorem.

Theorem 3.4. *The product of a bounded sequence by an infinitesimal sequence is an infinitesimal sequence.*

Proof. Let $\{x_n\}$ be a bounded sequence and $\{\alpha_n\}$ an infinitesimal sequence. Since the sequence $\{x_n\}$ is bounded, there exists a number $A > 0$ such that any term x_n satisfies the inequality $|x_n| \leqslant A$. Let us take an arbitrary positive number ε. Since the sequence $\{\alpha_n\}$ is infinitesimal, there must be given, for the positive number $\frac{\varepsilon}{A}$, a number N such that when $n \geqslant N$ there holds an inequality $|\alpha_n| < \frac{\varepsilon}{A}$. Then, for $n \geqslant N$ we have $|x_n \cdot \alpha_n| = |x_n| \cdot |\alpha_n| < A\frac{\varepsilon}{A} = \varepsilon$. Hence, the sequence $\{x_n \cdot \alpha_n\}$ is infinitesimal and this proves the theorem.

Corollary. *The product of any finite number of infinitesimal sequences is an infinitesimal sequence.*

Remark. The quotient of two infinitesimal sequences may be a sequence of any type and may even be senseless. If, for instance, $\alpha_n = \frac{1}{n}$, $\beta_n = \frac{1}{n}$, then all the terms of the sequence $\left\{\frac{\alpha_n}{\beta_n}\right\}$ are equal to unity. If $\alpha_n = \frac{1}{n}$, $\beta_n = \frac{1}{n^2}$, the sequence $\left\{\frac{\alpha_n}{\beta_n}\right\}$ is infinitely large, and, conversely, if $\alpha_n = \frac{1}{n^2}$

* Here and in what follows the notation $a = \max\{a_1, a_2, \ldots, a_n\}$ means that the number a is equal to the maximum number of the numbers $a_1, a_2, \ldots, a_n$.

and $\beta_n = \frac{1}{n}$, the sequence $\left\{\frac{\alpha_n}{\beta_n}\right\}$ is infinitesimal. If infinitely many terms of the sequence $\{\beta_n\}$ are equal to zero, the quotient $\left\{\frac{\alpha_n}{\beta_n}\right\}$ is meaningless.

Theorem 3.5. If all the terms of the infinitesimal sequence $\{\alpha_n\}$ are equal to one and the same number c, then $c = 0$.

Proof. Assume that $c \neq 0$. Let us put $\varepsilon = \frac{|c|}{2}$, $\varepsilon > 0$. Beginning with a number N corresponding to that ε there holds an inequality $|\alpha_n| < \varepsilon$. Since $\alpha_n = c$ and $\varepsilon = \frac{|c|}{2}$, the last inequality can be rewritten as $|c| < \frac{|c|}{2}$, whence we have $1 < \frac{1}{2}$. The contradiction obtained shows that the supposition $c \neq 0$ cannot be true. Hence $c = 0$, and that is the proof of the theorem.

In conclusion we wish to make a statement relating infinitely large and infinitesimal sequences.

Theorem 3.6. If $\{x_n\}$ is an infinitely large sequence then, beginning with a certain number n, the sequence $\left\{\frac{1}{x_n}\right\}$ is specified and is infinitesimal. If all the terms of the infinitesimal sequence $\{\alpha_n\}$ are not equal to zero, then the sequence $\left\{\frac{1}{\alpha_n}\right\}$ is infinitely large.

Proof. Let us note, first of all, that only a finite number of terms of an infinitely large sequence can be equal to zero. Indeed, from the definition of an infinitely large sequence it follows that for a given positive number A there can be indicated a number N^* beginning with which there holds an inequality $|x_n| > A$. This means that for $n \geqslant N^*$ all the terms x_n are non-zero, and therefore the sequence $\left\{\frac{1}{x_n}\right\}$ is meaningful if its terms are considered beginning with the number N^*. Let us now prove that $\left\{\frac{1}{x_n}\right\}$ is an infinitesimal sequence. Let ε be an arbitrary positive number. We can indicate, for $\frac{1}{\varepsilon}$, a number $N \geqslant N^*$ such that at $n \geqslant N$ the terms x_n of the sequence $\{x_n\}$ satisfy the inequality $|x_n| > \frac{1}{\varepsilon}$. Therefore, the inequality $\left|\frac{1}{x_n}\right| < \varepsilon$ if fulfilled beginning with the number N we have indicated. Thus, we have proved that the sequence $\left\{\frac{1}{x_n}\right\}$ is infinitesimal.

The proof of the second part of the theorem can be carried out in quite a similar way.

3.2. CONVERGENT SEQUENCES AND THEIR BASIC PROPERTIES

3.2.1. The concept of a convergent sequence.

Definition. *The sequence* $\{x_n\}$ *is called* **convergent** *if there exists a number* a *such that the sequence* $\{x_n - a\}$ *is infinitesimal. In this case the number* a *is referred to as the* **limit of the sequence** $\{x_n\}$.*

Evidently, the definition of a convergent sequence can also be formulated as follows.

The sequence $\{x_n\}$ *is called convergent if there exists a number* a *such that for any positive number* ε *there can be chosen a number* N** *such that when* $n \geqslant N$ *all the terms* x_n *of that sequence satisfy the inequality*

$$|x_n - a| < \varepsilon. \tag{3.4}$$

In this case the number a *is referred to as the limit of a sequence.*

If the sequence $\{x_n\}$ converges and has the number a as its limit, it can be symbolized as follows***

$$\lim_{n\to\infty} x_n = a \text{ or } x_n \to a \text{ for } n \to \infty.$$

Remark 1. Inequality (3.4) is equivalent to the inequalities $-\varepsilon < x_n - a < +\varepsilon$, or $a - \varepsilon < x_n < a + \varepsilon$. The last inequalities imply that the term x_n is in the ε-neighbourhood of the number a (recall that the ε-neighbourhood of the number a is the interval $(a - \varepsilon, a + \varepsilon)$). Therefore, the definition of a convergent sequence can be also formulated as follows.

The sequence $\{x_n\}$ *is called convergent if there exists a number* a *such that any* ε*-neighbourhood of the number* a *contains all the terms of the sequence* $\{x_n\}$ *beginning with a certain number*****.

The definition of a convergent sequence states that the difference $x_n - a = \alpha_n$ is infinitesimal. Consequently, every term x_n of a convergent sequence having a as its limit can be represented in the

---

* In accordance with this definition every infinitesimal sequence is convergent and has the number zero for its limit.

** Since N depends on ε, it is sometimes written as $N = N(\varepsilon)$.

*** Note that infinitely large sequences are sometimes said to be converging to infinity. Therefore, if the sequence $\{x_n\}$ is infinitely large, it can be symbolized as follows:

$$\lim_{n\to\infty} x_n = \infty,$$

If the terms of an infinitely large sequence have a definite sign, beginning with a certain number, the sequence $\{x_n\}$ is said to be converging to the infinity of a definite sign. The notation is:

$$\lim_{n\to\infty} x_n = +\infty, \quad \lim_{n\to\infty} x_n = -\infty.$$

**** Surely, dependent on ε.

form

$$x_n = a + \alpha_n, \tag{3.5}$$

where α_n is a term of an infinitesimal sequence.

Remark 2. It is evident from the definition of the limit of a sequence that the finite number of terms does not affect the convergence of that sequence or the value of its limit.

Let us consider some examples of convergent sequences.

(1) The sequence $\left\{\frac{n}{n+1}\right\}$ is convergent; the limit of this sequence is equal to unity. In fact, since $\frac{n}{n+1}-1=-\frac{1}{n+1}$, it follows that in order to prove the assertion it is sufficient to ascertain that the sequence $\left\{-\frac{1}{n+1}\right\}$ is infinitesimal. If $n \geqslant N$, then $\left|-\frac{1}{n+1}\right| \leqslant \frac{1}{N+1}$, and therefore, given a positive ε, a requisite number N must be chosen from the condition $\frac{1}{N+1} < \varepsilon$ or $N > \frac{1}{\varepsilon} - 1$. For instance, it can be assumed that

$$N = \begin{cases} \left[\frac{1}{\varepsilon} - 1\right] + 1 & \text{for } \varepsilon \leqslant 1, \\ 1 & \text{for } \varepsilon > 1. \end{cases}$$

(2) Let us prove that the sequence $x_1 = 0.3$; $x_2 = 0.33$; . . .; $x_n = 0.\underbrace{33 \ldots 3}_{n \text{ times}}$; . . . is convergent and has the number 1/3 as its limit. Since the number 1/3 is representable as an infinite decimal fraction 0.333 . . ., the rule of comparison of real numbers (see 2.1.3) yields inequalities*

$$0.\underbrace{33 \ldots 3}_{n \text{ times}} \leqslant \frac{1}{3} \leqslant \underbrace{0.33 \ldots 3}_{n \text{ times}} + \frac{1}{10^n}.$$

We obtain from these inequalities that $\left|x_n - \frac{1}{3}\right| < \frac{1}{10^n}$. Since $\frac{1}{10^n} \leqslant \frac{1}{10^N}$ for $n \geqslant N$, we shall obtain $\left|x_n - \frac{1}{3}\right|$ for $n \geqslant N$ choosing, given an arbitrary positive number ε, a requisite number N from the condition $\frac{1}{10^N} < \varepsilon$.

The possibility of choosing a number N satisfying the condition $|q|^N < \varepsilon$ for any $|q| < 1$ was established in Example 1 in 3.1.3.

3.2.2. Basic properties of convergent sequences.

Theorem 3.7. *A convergent sequence has only one limit.*

---

* See also inequalities (2.5) in 2.1.4.

Proof. Let a and b be the limits of the convergent sequence $\{x_n\}$. Then, making use of the special representation (3.5) of the terms x_n of the convergent sequence $\{x_n\}$, we shall obtain $x_n = a + \alpha_n$, $x_n = b + \beta_n$, where α_n and β_n are the terms of the infinitesimal sequences $\{\alpha_n\}$ and $\{\beta_n\}$.

Subtracting one of the relations written above from the other, we find that $\alpha_n - \beta_n = b - a$. Since all the terms of the infinitesimal sequence $\{\alpha_n - \beta_n\}$ possess one and the same constant value $b - a$, we have, in accordance with Theorem 3.5, that $b - a = 0$, i.e. $b = a$. We have proved the theorem.

Theorem 3.8. *A convergent sequence is bounded.*

Proof. Let $\{x_n\}$ be a convergent sequence and a its limit. Making use of formula (3.5), we find that

$$x_n = a + \alpha_n,$$

where α_n is a term of an infinitesimal sequence. Since the infinitesimal sequence $\{\alpha_n\}$ is bounded (see Theorem 3.3), there is a number A such that the inequality $\{\alpha_n\} \leqslant A$ holds for all the numbers n. Therefore, $| x_n | \leqslant | a | + A$ for all the numbers n, which means that the sequence $\{x_n\}$ is bounded. That was the proof of the theorem

Remark 1. A bounded sequence is not necessarily convergent. For example, the sequence 1, —1, 1, —1, . . . is bounded but not convergent. Indeed, if that sequence converged to some number a, then each of the sequences $\{x_n - a\}$ and $\{x_{n+1} - a\}$ would be infinitesimal. But then, by virtue of Theorem 3.2, the sequence $\{(x_n - a) - (x_{n+1} - a)\} = \{x_n - x_{n+1}\}$ would be infinitesimal. But that is impossible since $| x_n - x_{n+1} | = 2$ for any number n.

Let us prove the following *basic* theorems.

Theorem 3.9. *The sum of the convergent sequences $\{x_n\}$ and $\{y_n\}$ is a convergent sequence whose limit is equal to the sum of the limits of the sequences $\{x_n\}$ and $\{y_n\}$.*

Proof. Let a and b be the limits of the sequences $\{x_n\}$ and $\{y_n\}$, respectively. Then

$$x_n = a + \alpha_n, \quad y_n = b + \beta_n,$$

where $\{\alpha_n\}$ and $\{\beta_n\}$ are infinitesimal sequences. Consequently, $(x_n + y_n) - (a + b) = \alpha_n + \beta_n$.

Thus, the sequence $\{(x_n + y_n) - (a + b)\}$ is infinitesimal and, therefore, the sequence $\{x_n + y_n\}$ is convergent and its limit is $a + b$.

Theorem 3.10. *The difference between the convergent sequences $\{x_n\}$ and $\{y_n\}$ is a convergent sequence whose limit is equal to the difference between the limits of the sequences $\{x_n\}$ and $\{y_n\}$.* The proof of this theorem is analogous to that of Theorem 3.9.

Theorem 3.11. *The product of the convergent sequences* $\{x_n\}$ *and* $\{y_n\}$ *is a convergent sequence whose limit is equal to the product of the limits of the sequences* $\{x_n\}$ *and* $\{y_n\}$.

Proof. If a and b are the limits of the sequences $\{x_n\}$ and $\{y_n\}$, respectively, then $x_n = a + \alpha_n$, $y_n = b + \beta_n$ and $x_n \cdot y_n = a \cdot b + a \cdot \beta_n + b \cdot \alpha_n + \alpha_n \cdot \beta_n$. Consequently

$$x_n \cdot y_n - a \cdot b = a \cdot \beta_n + b \cdot \alpha_n + \alpha_n \cdot \beta_n.$$

In accordance with Theorem 3.4 and its corollary and also with Theorem 3.1, the sequence $\{a \cdot \beta_n + b \cdot \alpha_n + a_n \cdot \beta_n\}$ is infinitesimal, that is, the sequence $\{x_n \cdot y_n - a \cdot b\}$ is also infinitesimal and, therefore, the sequence $\{x_n \cdot y_n\}$ is convergent and the number $a \cdot b$ is its limit.

To prove the corresponding theorem for the quotient of two sequences, we shall need the following lemma.

Lemma 1. *If the sequence* $\{y_n\}$ *is convergent and possesses a non-zero limit* b, *then, beginning with a certain number, the sequence* $\left\{\frac{1}{y_n}\right\}$ *is defined and proves to be bounded.*

Proof. Assume $\varepsilon = \frac{|b|}{2}$. Since $b \neq 0$, $\varepsilon > 0$. Let N be the number corresponding to that ε beginning with which there holds an inequality $|y_n - b| < \varepsilon$ or $|y_n - b| < \frac{|b|}{2}$. It follows from this inequality that for $n \geqslant N$ there holds an inequality* $|y_n| > \frac{|b|}{2}$. Therefore, given $n \geqslant N$, we have $\left|\frac{1}{y_n}\right| > \frac{2}{|b|}$. Consequently, beginning with that number N, we can consider the sequence $\left\{\frac{1}{y_n}\right\}$, and this sequence is bounded. We have proved the lemma.

Theorem 3.12. *The quotient of two convergent sequences* $\{x_n\}$ *and* $\{y_n\}$, *under the condition that the limit of* $\{y_n\}$ *is non-zero, is a convergent sequence whose limit is equal to the quotient of the limits of the sequences* $\{x_n\}$ *and* $\{y_n\}$.

Proof. It follows from Lemma 1 proved above, that beginning with a certain number N, the terms of the sequence $\{y_n\}$ are different from zero and that the sequence $\left\{\frac{1}{y_n}\right\}$ is bounded. We shall consider the sequence $\left\{\frac{x_n}{y_n}\right\}$ beginning with that particular number. Let a and b be the limits of the sequences $\{x_n\}$ and $\{y_n\}$,

---

* Indeed, since $b = (b - y_n) + y_n$ and $|b - y_n| < \frac{|b|}{2}$, we have $|b| \leqslant |b - y_n| + |y_n| < \frac{|b|}{2} + |y_n|$.

respectively. Let us prove that the sequence $\left\{\frac{x_n}{y_n}-\frac{a}{b}\right\}$ is infinitesimal. Indeed, since $x_n=a+\alpha_n$, $y_n=b+\beta_n$, we have

$$\frac{x_n}{y_n}-\frac{a}{b}=\frac{x_n\cdot b-y_n\cdot a}{y_n\cdot b}=\frac{1}{y_n}\left(\alpha_n-\frac{a}{b}\beta_n\right).$$

Since the sequence $\left\{\frac{1}{y_n}\right\}$ is bounded and the sequence $\left\{\alpha_n-\frac{a}{b}\beta_n\right\}$ is infinitesimal, it follows that the sequence $\left\{\frac{1}{y_n}\cdot\left(\alpha_n-\frac{a}{b}\beta_n\right)\right\}\equiv\left\{\frac{x_n}{y_n}-\frac{a}{b}\right\}$ is infinitesimal. That is the proof of the theorem.

3.2.3. Limiting process in inequalities. We have just found that arithmetic operations on convergent sequences lead to the same operations on their limits. Here we will show that the inequalities satisfied by the terms of convergent sequences, pass, in the limit, into the corresponding inequalities for the limits of these sequences.

Theorem 3.13. *If, beginning with a certain number, the terms of a convergent sequence* $\{x_n\}$ *satisfy the inequality* $x_n\geqslant b$ $(x_n\leqslant b)$, *then the limit* a *of that sequence, too, satisfies the inequality* $a\geqslant b$ $(a\leqslant b)$.

Proof. Let all the terms x_n, at least beginning with a certain number, satisfy the inequality $x_n\geqslant b$. It is required to prove the inequality $a\geqslant b$. Suppose that $a<b$. Since a is the limit of the sequence $\{x_n\}$, there can be, for a positive $\varepsilon=b-a$, a number N such that when $n\geqslant N$, there holds an inequality $|x_n-a|<b-a$. This inequality is equivalent to the following two inequalities: $-(b-a)<x_n-a<b-a$. Making use of the right-hand inequality, we obtain $x_n<b$, but this contradicts the condition of the theorem. The case $x_n\leqslant b$ is treated analogously. The theorem has been proved.

Remark. The terms of the convergent sequence $\{x_n\}$ can satisfy the strict inequality $x_n>b$. However, in this case the limit a may turn out to be equal to b. For instance, if $x_n=\frac{1}{n}$, then $x_n>0$, but $\lim\limits_{n\to\infty}x_n=0$.

Corollary 1. *If, beginning with a certain number, the terms* x_n *and* y_n *of the convergent sequences* $\{x_n\}$ *and* $\{y_n\}$ *satisfy the inequality* $x_n\leqslant y_n$, *then their limits comply with a similar inequality.*

$$\lim_{n\to\infty}x_n\leqslant\lim_{n\to\infty}y_n.$$

Indeed, the terms of the sequence $\{y_n-x_n\}$ being nonnegative, its limit $\lim\limits_{n\to\infty}(y_n-x_n)=\lim\limits_{n\to\infty}y_n-\lim\limits_{n\to\infty}x_n$ is nonnegative too. Hence it follows that

$$\lim_{n\to\infty}x_n\leqslant\lim_{n\to\infty}y_n.$$

Corollary 2. *If all the terms of the convergent sequence* $\{x_n\}$ *are contained in the closed interval* $[a, b]$, *its limit* c *also belongs to that interval.*

Indeed, since $a \leqslant x_n \leqslant b$, it follows that $a \leqslant c \leqslant b$.

The following theorem plays an important role in various applications.

Theorem 3.14. *Let* $\{x_n\}$ *and* $\{z_n\}$ *be convergent sequences having a common limit* a. *Suppose, in addition, that beginning with a certain number the terms of the sequence* $\{y_n\}$ *satisfy the inequalities* $x_n \leqslant y_n \leqslant z_n$. *Then the sequence* $\{y_n\}$ *converges and has a limit* a.

Proof. It is sufficient to prove that the sequence $\{y_n - a\}$ is infinitesimal. Let us denote by N^* the number beginning with which the inequalities indicated in the hypothesis are satisfied. Then, beginning with the same number, the inequalities $x_n - a \leqslant y_n - a \leqslant z_n - a$ will be satisfied as well. Hence it follows that for $n \geqslant N^*$ the terms of the sequence $\{y_n - a\}$ satisfy the inequality

$$| y_n - a | \leqslant \max \{| x_n - a |, | z_n - a |\}.$$

Since $\lim\limits_{n\to\infty} x_n = a$ and $\lim\limits_{n\to\infty} z_n = a$, there can be indicated, for any $\varepsilon > 0$, the numbers N_1 and N_2 such that $| x_n - a | < \varepsilon$ for $n \geqslant N_1$ and $| z_n - a | < \varepsilon$ for $n \geqslant N_2$. Suppose $N = \max \{N^*, N_1, N_2\}$. Beginning with this number, there holds an inequality $| y_n - a | < \varepsilon$. Thus we see that the sequence $\{y_n - a\}$ is infinitesimal. That is the proof of the theorem.

3.3. MONOTONIC SEQUENCES

3.3.1. Definition of monotonic sequences.

Definition. *The sequence* $\{x_n\}$ *is said to be* **nondecreasing (nonincreasing)** *if every following term of that sequence is not smaller (not larger) than the preceding term, i.e. if for all the numbers* n *there holds an inequality*

$$x_n \leqslant x_{n+1} \quad (x_n \geqslant x_{n+1}).$$

Both nondecreasing and nonincreasing sequences are referred to as *monotonic (monotone) sequences.* If the terms of the monotone sequence $\{x_n\}$ satisfy the inequality $x_n < x_{n+1}$ $(x_n > x_{n+1})$ for all the numbers n, the sequence $\{x_n\}$ is called *monotone increasing (monotone decreasing).* Monotone increasing and monotone decreasing sequences are also called *strictly monotonic.*

Monotonic sequences have either an upper or a lower bound, namely, *nonincreasing* sequences are *bounded above* and *nondecreasing* sequences are *bounded below* by their first terms. Therefore, a nonincreasing sequence has both an upper bound and a lower bound if it

is bounded below and a nondecreasing sequence has both an upper bound and a lower bound if it is bounded above.

Here are some examples of monotonic sequences.

1. The sequence $1, 1, \frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{n}, \frac{1}{n}, \ldots$ is nonincreasing. It has its first term equal to unity as its upper bound and the number zero as its lower bound.

2. The sequence $1, 1, 2, 2, \ldots, n, n, \ldots$ is nondecreasing. It has its first term equal to unity as its lower bound and has no upper bound.

3. The sequence $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \ldots \frac{n}{n+1}, \ldots$ is an increasing sequence. It is bounded above and below, having its first term $\frac{1}{2}$ as its lower bound and the number, say, unity, as its upper bound.

3.3.2. Test for convergence of a monotonic sequence.

Here we must quote the following *fundamental* theorem.

Theorem 3.15. *If a nondecreasing (nonincreasing) sequence $\{x_n\}$ is bounded above (below), it is convergent.*

In accordance with what was said in the previous subsection, the sequence $\{x_n\}$ which satisfies the condition of Theorem 3.15 is bounded. Therefore, Theorem 3.15 can be briefly formulated as follows: *if the monotonic sequence $\{x_n\}$ has both an upper bound and a lower bound, it is convergent.*

Proof. Since the sequence $\{x_n\}$ is bounded, the set of its terms has the supremum $\bar{x}$ and the infimum $\underline{x}$ (see Theorem 2.1). Let us prove that if $\{x_n\}$ is a nondecreasing sequence, then its limit is the indicated supremum $\bar{x}$, now if $\{x_n\}$ is a nonincreasing sequence, then its limit is the indicated infimum $\underline{x}$. We shall restrict our considerations to a nondecreasing sequence since the reasoning for a nonincreasing sequence is quite analogous.

Inasmuch as $\bar{x}$ is the least upper bound of the set of terms of the sequence $\{x_n\}$, there is a term x_N, for any $\varepsilon > 0$, such that $x_N > \bar{x} - \varepsilon$ and $x_N \leqslant \bar{x}$ (any term x_n is not larger than the supremum $\bar{x}$, $x_n \leqslant \bar{x}$). Comparing the indicated inequalities, we obtain inequalities $0 \leqslant \bar{x} - x_N < \varepsilon$. Since $\{x_n\}$ is a nondecreasing sequence, it follows that for $n \geqslant N$ there hold inequalities $x_N \leqslant x_n \leqslant \bar{x}$. And this implies that for $n \geqslant N$ there hold inequalities $0 \leqslant \bar{x} - x_n \leqslant \bar{x} - x_N$. As was noted above $\bar{x} - x_N < \varepsilon$, consequently, for $n \geqslant N$, there hold inequalities $0 \leqslant \bar{x} - x_n < \varepsilon$, which yield an inequality $|x_n - \bar{x}| < \varepsilon$. Hence, we have established that $\bar{x}$ is the limit of the sequence $\{x_n\}$ and have proved the theorem.

Remark 1. *The condition of boundedness of a monotonic sequence is a necessary and sufficient condition for its convergence.*

Indeed, if a monotone sequence is bounded, then it converges, in accordance with Theorem 3.15; and if a monotone sequence converges, then, by virtue of Theorem 3.8, it is bounded.

Remark 2. A convergent sequence is not necessarily monotonic. An example is a sequence $\{x_n\}$ for which $x_n = \frac{(-1)^n}{n}$ converges and has the number zero as its limit. As the signs of the terms of this sequence alternate, it is not monotonic.

Remark 3. If the sequence $\{x_n\}$ is nondecreasing and bounded, and $\bar{x}$ is its limit, then for all the numbers n there holds an inequality $x_n \leqslant \bar{x}$. The terms of the nonincreasing bounded sequence $\{x_n\}$, converging to $\underline{x}$, satisfy the inequality $\underline{x} \leqslant x_n$. The validity of this assertion was established in the process of proving Theorem 3.15.

Corollary of Theorem 3.15. *Suppose there is an infinite system of closed intervals* $[a_1, b_1]$, $[a_2, b_2]$, $[a_3, b_3]$, . . ., $[a_n, b_n]$, . . ., *each succeeding interval of which is contained in the preceding one,* and suppose the difference* $b_n - a_n$ *(we shall call it the length of the interval* $[a_n, b_n]$*) tends to zero as* $n \to \infty$ *(the system of intervals possessing this property will be referred to as a nested system). Then, there exists a point* c *belonging to all the intervals of the system, and that point is unique.*

Proof. We must note first of all that there can be only one point c belonging to all the closed intervals. Indeed, if there could be found some other point d belonging to all the intervals, then the whole interval** $[c, d]$ would belong to all the intervals $[a_n, b_n]$. But then, for any number n, there would hold the inequality $b_n - a_n \geqslant d - c > 0$, which is impossible since $b_n - a_n \to 0$ as $n \to \infty$. Now we shall prove that there *exists* a point c belonging to all the intervals $[a_n, b_n]$. Since the system of intervals is nested, the sequence of the left-hand end-points of $\{a_n\}$ is nondecreasing and the sequence of the right-hand end-points of $\{b_n\}$ is nonincreasing. Since both these sequences are bounded (all the terms of the sequences $\{a_n\}$ and $\{b_n\}$ belong to the interval $[a_1, b_1]$), by virtue of Theorem 3.15 they are both convergent. The difference $b_n - a_n$ being infinitesimal, the indicated sequences have a common limit which we shall denote by c. It follows from Remark 3 that for any number n the inequalities $a_n \leqslant c \leqslant b_n$ hold, i.e. point c belongs to all intervals $[a_n, b_n]$.

3.3.3. Some examples of convergent monotonic sequences. Let us consider some sequences whose limit can be found with the aid of Theorem 3.15 on the limit of a monotonic sequence. Besides, here we shall get acquainted with a general method of finding limits of sequences specified by recurrence formulas***.

* This means that $a_{n-1} \leqslant a_n \leqslant b_n \leqslant b_{n-1}$.

** For definiteness we assume that $d > c$.

*** A recurrence formula (from the Latin word recurrens meaning returning) is a formula making it possible to express the $(n+1)$th term of a sequence through the values of its first n terms.

Example 1. Let us consider a sequence $\{x_n\}$ whose term x_n is equal to

$$x_n = \sqrt{a+\sqrt{a+\sqrt{a+\ldots+\sqrt{a}}}}, \quad a>0.$$

This sequence can, evidently, be specified by the following recurrence formula:

$$x_1 = \sqrt{a}, \quad x_n = \sqrt{a + x_{n-1}}.$$

To ascertain that the sequence $\{x_n\}$ has a limit, we shall prove that it is *increasing* and *bounded*. The first fact does not require verification. We shall prove that the sequence $\{x_n\}$ is bounded above by the number A, where A is the greater of the two numbers a and 2. If $x_n \leqslant a$, the proof is complete. If, now, $x_n > a$, then, replacing, in the right-hand side of the inequality $x_n^2 = a + x_{n-1} \leqslant a + x_n$, the number a by the number x_n exceeding it, we shall obtain $x_n^2 < 2x_n$, whence $x_n < 2$. Thus we have proved that the sequence $\{x_n\}$ is bounded above. In accordance with Theorem 3.15 it has a limit, which will be denoted by c. Evidently, $c > 0$. The recurrence formula yields a relation

$$x_n^2 = a + x_{n-1},$$

which means that the sequences $\{x_n^2\}$ and $\{a + x_{n-1}\}$ are identical. Therefore, their limits are equal. Since the first of these sequences has a limit c^2, and the second $a + c$, we have a relation $c^2 = a + c$, whence we can find, since c is positive, that $c = \frac{1+\sqrt{1+4a}}{2}$.

Example 2. Let us now consider the sequence $\{x_n\}$ which is usually used to compute a square root of a positive number a on modern computers. This sequence is specified by the following recurrence formula:

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right), \quad n = 1, 2, \ldots,$$

where x_1 can be any positive number.

Let us prove that this sequence converges and has the number $\sqrt{a}$ as its limit. First of all, we shall prove that the sequence $\{x_n\}$ has a limit. To prove this fact, it suffices to establish that the sequence $\{x_n\}$ is *bounded below*, and *is nonincreasing beginning with the second number*. First we shall prove that the sequence $\{x_n\}$ is bounded below. By the hypothesis $x_1 > 0$. But then, from the recurrence formula taken for $n = 1$, it follows that $x_2 > 0$. From this inequality and from the same formula taken for $n = 2$ it follows that $x_3 > 0$. Proceeding with our reasoning, we shall prove that all x_n are positive.

Let us now prove that *for* $n \geqslant 2$ *all the* x_n *satis fy the inequality* $x_n \geqslant \sqrt{a}$. Having rewritten the recurrence formula in the form

$x_{n+1}=\frac{\sqrt{a}}{2}\left(\frac{x_n}{\sqrt{a}}+\frac{\sqrt{a}}{x_n}\right)$, we shall make use of an almost obvious inequality $t+\frac{1}{t}\geqslant 2$*, valid for any $t>0$ $\left(\text{we take } t=\frac{x_n}{\sqrt{a}}\right)$. We shall see that $x_{n+1}\geqslant\sqrt{a}$ for any $n\geqslant 1$, i.e. $x_n\geqslant\sqrt{a}$ beginning with the number $n=2$.

The last fact we have to prove is that the *sequence* $\{x_n\}$ *is nonincreasing for* $n\geqslant 2$. We obtain from the recurrence formula that $\frac{x_{n+1}}{x_n}=\frac{1}{2}\left(1+\frac{a}{x_n^2}\right)$, whence, taking into account that $x_n\geqslant\sqrt{a}$, we shall find $\frac{x_{n+1}}{x_n}\leqslant 1$, or $x_n\geqslant x_{n+1}$ (for $n\geqslant 2$).

Since the sequence $\{x_n\}$ is nonincreasing for $n\geqslant 2$ and is bounded below by the number $\sqrt{a}$, it has a limit not smaller than $\sqrt{a}$ (see Theorem 3.15 and Theorem 3.13). Denoting this limit by c and taking into account that $\lim\limits_{n\to\infty} x_{n+1}=c$ and $\lim\limits_{n\to\infty}\left\{\frac{1}{2}\left(x_n+\frac{a}{x_n}\right)\right\}=$ $=\frac{1}{2}\left(c+\frac{a}{c}\right)$ we obtain $c=\frac{1}{2}\left(c+\frac{a}{c}\right)$**. Hence, $c=\sqrt{a}$.

Remark 1. In the examples discussed above we made use of the following frequently employed method of finding the limit of a sequence. First the existence of the limit is established and then its numerical value is found from the equation which can be obtained from the recurrence formula if x_n and x_{n+1} entering into it are replaced by the sought-for value c of the limit of the sequence $\{x_n\}$.

Remark 2. Recurrence formulas are frequently used today since they imply multiple repetition of single-type operations, and that is especially convenient for computations carried out by means of high-speed computers.

As we have seen, the recurrence formula discussed above establishes the algorithm of calculating $\sqrt{a}$ (we have proved that $\lim\limits_{n\to\infty} x_n=\sqrt{a}$).

Supplement 2 to the present chapter discusses the rate of convergence of the sequence $\{x_n\}$ to $\sqrt{a}$. It is proved there that if, for any $a>1$, we make a proper choice of the first approximation x_1, we can obtain the number $\sqrt{a}$ with an error not exceeding 10^{-10} already in the fourth approximation x_4.

Example 3. Let us prove that the sequence $\{c_n\}$, for which $c_n=\frac{x^{n+1}}{(n+1)!}$, has, for any fixed x, a limit equal to zero. Since

* To prove this inequality, it suffices to note that for $t>0$ it is equivalent to the inequality $t^2-2t+1\geqslant 0$.

** This equality follows from the recurrence formula

$$x_{n+1}=\frac{1}{2}\left(x_n+\frac{a}{x_n}\right).$$

for any sufficiently large n the fraction $\frac{|x|}{n+1}$ is less than unity, we have, beginning with a certain number N, $|c_{n+1}|<|c_n|$ inasmuch as $|c_{n+1}|=\frac{|x|^n}{n!}\cdot\frac{|x|}{n+1}=|c_n|\cdot\frac{|x|}{n+1}$.

Consequently, beginning with the number N, the sequence $\{|c_n|\}$ is monotone decreasing and bounded below (by zero, for instance). In accordance with Theorem 3.15 the sequence $\{|c_n|\}$ converges. Let c be the limit of that sequence. The relation $|c_{n+1}|=|c_n|\cdot\frac{|x|}{n+1}$ yields $c=0$ since the limit of the sequence $\{|c_{n+1}|\}$ is equal to c and the limit of the sequence $\left\{\frac{|x|}{n+1}\right\}$ is equal to zero.

3.3.4. Number *e*. Let us apply Theorem 3.15 on the existence of a limit of a monotonic sequence to prove the existence of a limit of a sequence $\{x_n\}$ whose term x_n is determined by the equation

$$x_n=\left(1+\frac{1}{n}\right)^n.$$

We shall prove that the sequence *increases and is bounded above*. Applying Newton's binomial formula, we shall find that

$$x_n=1+n\frac{1}{n}+\frac{n(n-1)}{2!}\frac{1}{n^2}+\frac{n(n-1)(n-2)}{3!}\frac{1}{n^3}+\ldots+$$
$$+\frac{n(n-1)(n-2)\ldots[n-(n-1)]}{n!}\frac{1}{n^n}.$$

Let us write this expression in the following form:

$$x_n=2+\frac{1}{2!}\left(1-\frac{1}{n}\right)+\frac{1}{3!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)+\ldots$$
$$\ldots+\frac{1}{n!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{n-1}{n}\right). \qquad (3.6)$$

We shall use a similar form to write the term x_{n+1}:

$$x_{n+1}=2+\frac{1}{2!}\left(1-\frac{1}{n+1}\right)+\frac{1}{3!}\left(1-\frac{1}{n+1}\right)\left(1-\frac{2}{n+1}\right)+\ldots$$
$$\ldots+\frac{1}{(n+1)!}\left(1-\frac{1}{n+1}\right)\left(1-\frac{2}{n+1}\right)\ldots\left(1-\frac{n}{n+1}\right).$$

A direct comparison convincingly shows that*

$$x_n<x_{n+1},$$

or that the sequence $\{x_n\}$ is *increasing*.

---

* Because $\left(1-\frac{k}{n}\right)<\left(1-\frac{k}{n+1}\right)$ for any $0<k<n$ and, besides, x_{n+1} possesses an additional positive term as compared to x_n.

To prove that the sequence is bounded above, we make a note of the fact that each expression in the parentheses in relation (3.6) is less than unity. Taking into account the fact that $\frac{1}{k!} < \frac{2}{2^{k-1}}$ for $k \geqslant 2$, we obtain

$$x_n < 2 + \frac{1}{2} + \frac{1}{2^2} + \ldots + \frac{1}{2^{n-1}} = 3 - \frac{1}{2^{n-1}} < 3.$$

Thus we see that the sequence $\{x_n\}$ increases and is bounded above. In accordance with Theorem 3.15 the sequence $\{x_n\}$ has a limit. This limit is referred to as the number e. Consequently, by definition,

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n.$$

Remark. We shall see later on that the number e plays a significant role in mathematics. We have given here only the definition of the number e without indicating the method of computation of that number with any degree of accuracy. An algorithm of computing the number e will be given in 8.16.1 and 8.16.2.

For the time being we shall only take notice of the fact that since $x_n < 3$ and since it follows directly from (3.6) that $2 < x_n$, the number e has the limits

$$2 \leqslant e \leqslant 3 \tag{3.7}$$

(by virtue of Corollary 2 of Theorem 3.13).

3.4. SOME PROPERTIES OF ARBITRARY SEQUENCES AND NUMBER SETS

3.4.1. Subsequences of number sequences. Let $x_1, x_2, \ldots, x_n \ldots$ be a number sequence. We shall discuss an arbitrary increasing sequence of positive integers $k_1, k_2, \ldots, k_n, \ldots$. For our purpose we shall select from the sequence $\{x_n\}$ the terms with the numbers $k_1, k_2, \ldots, k_n \ldots$ and arrange them in the same order as the numbers k_n:

$$x_{k_1}, x_{k_2}, \ldots, x_{k_n}, \ldots.$$

This number sequence will be referred to as a *subsequence* of the sequence $\{x_n\}$. In a particular case, the sequence $\{x_n\}$ itself can be regarded as a subsequence (in that case $k_n = n$). We wish to note the following property of subsequences of a convergent sequence: *if the sequence $\{x_n\}$ converges and has the number a as its limit, then any subsequence of that sequence converges too and has a limit a.* Indeed, since $\{x_n\}$ is a convergent sequence and a is its limit, there exists, for any $\varepsilon > 0$, a number N such that when $n \geqslant N$, there holds an inequality $|x_n - a| < \varepsilon$. Let $\{x_{k_n}\}$ be a subsequence of the sequence $\{x_n\}$. Since $k_N \geqslant N$, it follows that beginning with the number k_N, the terms of the subsequence $\{x_{k_n}\}$ satisfy the inequality $|x_{k_n} - a| <$

$< \varepsilon$. Therefore, the subsequence $\{x_{k_n}\}$ converges and has the number a as its limit. The converse proposition is also true: *if all the subsequences of the given sequence $\{x_n\}$ converge, then the limits of all those subsequences are equal to one and the same number a; in particular, the sequence $\{x_n\}$ converges to that same number a.* In fact, since the sequence $\{x_n\}$ is also a subsequence, it converges and has, as its limit, a certain number a. But then any other subsequence also converges and has the same limit a.

The same property is inherent in the subsequences of infinitely large sequences, namely, *every subsequence of an infinitely large sequence is infinitely large as well.* The proof of this statement is analogous to that of the corresponding proposition concerning subsequences of convergent sequences.

Remark. Every convergent sequence contains a monotonic convergent subsequence. Indeed, if $\{x_n\}$ is a convergent sequence, and a is its limit, then at least one of the following three cases occurs: (1) there are infinitely many terms of the sequence which are equal to a, (2) there are infinitely many terms satisfying the inequality $x_n < a$ in any ε-neighbourhood of point a, (3) there are infinitely many terms satisfying the inequality $x_n > a$ in any ε-neighbourhood of point a.* In the first case, a subsequence of terms equal to a is a monotone convergent subsequence. The second and third cases are treated similarly, so we shall restrict the discussion to the second case, i.e. we shall assume that there are infinitely many terms x_n, in any ε-neighbourhood of point a, which satisfy the inequality $x_n < a$. To put it otherwise, we shall consider a case when any interval $(a - \varepsilon, a)$ contains an infinitude of terms of a sequence. Let x_{k_1} be one of those terms, $x_{k_1} < a$. From the infinitude of the terms of the sequence $\{x_n\}$ belonging to the interval (x_{k_1}, a), we shall choose some term x_{k_2}, whose number k_2 is greater than k_1. Then from the infinitude of terms of the sequence $\{x_n\}$ contained within the interval (x_{k_2}, a), we shall choose the term x_{k_3} for which $k_3 > k_2$. Proceeding in this way indefinitely, we shall obtain a monotone increasing subsequence $\{x_{k_n}\}$ of the sequence $\{x_n\}$ which converges to a by virtue of the property of subsequences of a convergent sequence indicated above in this subsection.

Note that a monotone infinitely large subsequence can be isolated from any infinitely large sequence.

3.4.2. Limit points of a sequence.

Definition 1. *Point x of an infinite number line is called a* **limit point (accumulation point)** *of a sequence $\{x_n\}$ if, in any ε-neighbourhood of that point, there are infinitely many terms of the sequence $\{x_n\}$.*

* If neither of these cases occurred, then a certain ε-neighbourhood of point a would contain only a finite number of terms of the sequence, that is, point a would not be the limit of the sequence.

We can assert that the following lemma is valid.

Lemma 2. *If x is a limit point of the sequence $\{x_n\}$, then it is possible to isolate from this sequence a subsequence $\{x_{k_n}\}$ converging to the number x.*

Proof. Let x be a limit point of the sequence $\{x_n\}$. We shall consider a system of ε-neighbourhoods of point x for which ε is successively equal to $1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}, \ldots$. In the first of these neighbourhoods we shall take a term x_{k_1} of the sequence $\{x_n\}$, in the second neighbourhood, a term x_{k_2} such that $k_2 > k_1$. In the third neighbourhood we shall take a term x_{k_3} such that $k_3 > k_2$. This process can be continued indefinitely since in any ε-neighbourhood of point x there are infinitely many terms of the sequence $\{x_n\}$. As a result, we shall obtain a subsequence $x_{k_1}, x_{k_2}, \ldots, x_{k_n}, \ldots$ of the sequence $\{x_n\}$ converging to x as $|x_{k_n} - x| < \frac{1}{n}$. The lemma has been proved.

Remark. The converse statement is also true: if it is possible to isolate from the sequence $\{x_n\}$ a subsequence converging to a number x, then the number x is a limit point of the sequence $\{x_n\}$. Indeed, in any ε-neighbourhood of point x there are infinitely many terms of the isolated subsequence and, hence, of the sequence $\{x_n\}$ itself.

Thus we can state another definition of the limit point of a sequence equivalent to Definition 1.

Definition 2. *Point x is called the limit point of the sequence $\{x_n\}$ in case a subsequence converging to x can be isolated from that sequence.*

Let us take a note of the following assertion.

Lemma 3. *Every convergent sequence has only one limit point coinciding with the limit of that sequence.*

Proof. We shall note first of all that the limit a of the convergent sequence $\{x_n\}$ is the limit point of that sequence since any ε-neighbourhood of point a contains all the terms of the sequence beginning with a certain number. Let us make sure that the convergent sequence has no other limit points. In fact, assume that b is the limit point of the convergent sequence. In accordance with Lemma 2, we can isolate from $\{x_n\}$ a subsequence $\{x_{k_n}\}$ converging to b, but any subsequence of a convergent sequence has the limit a (see 3.4.1), and therefore $b = a$.

Let us consider an example of a sequence having two limit points. We shall prove that the sequence

$$1, 2, \frac{1}{2}, 2, \frac{1}{3}, 2, \ldots, \frac{1}{n}, 2, \ldots$$

has only two limit points 0 and 2. It is evident that these points are the limit points of the sequence being discussed since the subsequence $1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}, \ldots$ of that sequence possesses the limit zero,

and the subsequence 2, 2, . . ., 2, . . . possesses the limit 2,* and has no other limit points. Indeed, let x be any point of the number scale different from the points 0 and 2. Let us consider the non-overlapping ε-neighbourhoods of the points 0, 2 and x (Fig. 3.1). The ε-neighbourhoods of the points 0 and 2 contain, beginning with

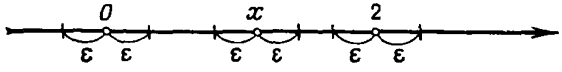


Fig. 3.1

a certain number, all the terms of the sequence and therefore the indicated ε-neighbourhood of point x contains only a finite number of its terms, that is x is not a limit point.

3.4.3. Existence of a limit point of a bounded sequence. The following remarkable assertion is valid.

Theorem 3.16. *Every bounded sequence possesses at least one limit point.*

Proof. Inasmuch as the sequence $\{x_n\}$ is bounded, there exist real numbers m and M such that all the terms x_n of the sequence $\{x_n\}$

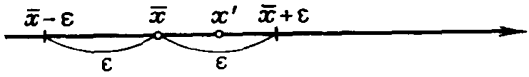


Fig. 3.2

satisfy the inequalities $m \leqslant x_n \leqslant M$. *Let us consider a set $\{x\}$ of real numbers x such that to the right** of each of these numbers there are either no terms of the sequence $\{x_n\}$ at all or the number of such terms is finite.* The set $\{x\}$ has at least one such term (say, number M) and is bounded below (by any number smaller than m). By virtue of Theorem 2.1 the set $\{x\}$ has an *infimum* which will be denoted by $\bar{x}$***.

We shall now prove that the number $\bar{x}$ is the limit point of the sequence $\{x_n\}$. Let ε be any positive number. It is obvious that the number $\bar{x} - \varepsilon$ does not belong to the set $\{x\}$, and, therefore, *to the right of the number $\bar{x} - \varepsilon$ there are infinitely many terms of the sequence $\{x_n\}$*. By the definition of an infimum, there is a number x' of the set $\{x\}$ satisfying the inequalities $\bar{x} \leqslant x' < \bar{x} + \varepsilon$ (Fig. 3.2). In accordance with the definition of the set $\{x\}$, *to the right of x' there*

* See Definition 2 of a limit point.

** We say that the number a is to the right of the number b if $a > b$ (see 2.3).

*** Later on we shall explain why it is expedient to denote the infimum as $\bar{x}$.

are not more than a finite number of terms of the sequence $\{x_n\}$. Consequently, in the half-open interval $(\bar{x} - \varepsilon, x']$, and, moreover, in the ε-neighbourhood of point $\bar{x}$ there are infinitely many terms of the sequence, i.e. $\bar{x}$ is the limit point of the sequence $\{x_n\}$. The proof has thus been completed.

Remak 1. Let us turn once again to the set $\{x\}$ that we introduced while proving Theorem 3.16. We have proved that the greatest lower bound $\bar{x}$ of that set is the limit point of the sequence $\{x_n\}$.

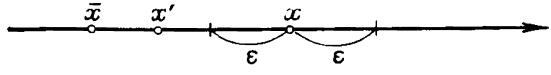


Fig. 3.3

We shall prove that *neither of the numbers* x *exceeding* $\bar{x}$ *is the limit point of the sequence* $\{x_n\}$, i.e. $\bar{x}$ is the *largest* limit point of that sequence. Let x be any number exceeding $\bar{x}$. We shall take a positive ε small enough for the number $x - \varepsilon$ to exceed $\bar{x}$ as well (Fig. 3.3). In accordance with the definition of the greatest lower bound there is a number x' from the set $\{x\}$ lying to the left of $x - \varepsilon$. The definition of the set $\{x\}$ states that to the right of x' and, therefore, in the ε-neighbourhood of point x there are not more than a finite number of terms of the sequence $\{x_n\}$. This proves that point x is not the limit point.

Definition. *The largest limit point* $\bar{x}$ *of the sequence* $\{x_n\}$ *is called the* **limit superior** *of that sequence and is symbolized as* $\bar{x} = \overline{\lim_{n\to\infty}} x_n$.

Remark 1 allows us to state that *every bounded sequence possesses a limit superior.*

The concept of the *limit inferior* $\underline{x}$ of the sequence $\{x_n\}$ is quite analogous and is defined as the least of the limits of that sequence. The limit inferior is designated as $\underline{x} = \underline{\lim_{n\to\infty}} x_n$.

The existence of a limit inferior of any bounded sequence $\{x_n\}$ is proved by a complete analogy with the argument used in Theorem 3.16 and Remark 1 to that theorem. Only this time we have to consider the set $\{x\}$ of real numbers x such that *to the left* of each of these numbers there are not more than a finite number of terms of this sequence.

Thus, we have come to the following assertion.

Every bounded sequence has both the limit superior and the limit inferior.

Here are two more corollaries that can be derived from Theorem 3.16 and Remark 1.

Corollary 1. *If* (a, b) *is an interval outside of which there are only a finite number of terms of the bounded sequence* $\{x_n\}$ *and* $\underline{x}$ *and* $\bar{x}$ *are the*

limit inferior and the limit superior of that sequence respectively, then the interval $(\underline{x}, \overline{x})$ *is contained within the interval* (a, b) *and therefore* $\overline{x} - \underline{x} \leqslant b - a$.

Proof. Since not more than a finite number of terms of the sequence are to the right of point b, point b belongs to the set $\{x\}$ indicated in the proof of Theorem 3.16 and, consequently, $\overline{x} \leqslant b$. By analogy we can prove that $a \leqslant \underline{x}$. And this means that the interval (a, b) contains the interval $(\underline{x}, \overline{x})$.

Corollary 2. *For any positive* ε *the interval* $(\underline{x} - \varepsilon, \overline{x} + \varepsilon)$ *contains all the terms of the sequence* $\{x_n\}$ *a certain number onward (surely, dependent on* ε*).*

Proof. Since $\overline{x}$ is the greatest lower bound of the set $\{x\}$ indicated in the proof of Theorem 3.16, there can be found, for any positive ε, a number x' smaller than $\overline{x} + \varepsilon$ and belonging to $\{x\}$. But this fact implies that to the right of x' and, therefore, *to the right of the interval* $(\underline{x} - \varepsilon, \overline{x} + \varepsilon)$ *there can be only a finite number of terms of the sequence* $\{x_n\}$. It can be proved by analogy that to the left of the interval $(\underline{x} - \varepsilon, \overline{x} + \varepsilon)$ there can be only a finite number of terms of the sequence $\{x_n\}$ either.

Remark 2. Let us now find out how many limit points a bounded sequence $\{x_n\}$ can possess.

We shall designate by $\underline{x}$ and $\overline{x}$, respectively, the limit inferior and the limit superior of that sequence. Evidently, all the limit points of the sequence $\{x_n\}$ (whatever their number) lie on the closed interval $[\underline{x}, \overline{x}]$.

If $\underline{x} = \overline{x}$*, the sequence has only one limit point. Now if $\underline{x} < \overline{x}$, the sequence has at least two limit points $\underline{x}$ and $\overline{x}$. Note that a sequence can have any number and even infinitely many limit points. The sequence $1, 2, \frac{1}{2}, 2, \ldots, \frac{1}{n}, 2, \ldots$ discussed in the preceding subsection has only two limit points, the limit inferior $\underline{x} = 0$ and the limit superior $\overline{x} = 2$. Now we shall cite an example of a sequence possessing infinitely many limit points. Let us consider, for example, a sequence whose terms run through, without recurrence, all the rational numbers of the closed interval $[0, 1]$**.

* We shall prove later on that the equality $\underline{x} = \overline{x}$ and the condition of boundedness are the necessary and sufficient conditions of the convergence of a sequence.

** The rational numbers of the interval $[0, 1]$ can be arranged in a sequence without recurrence in the following way, for instance. Let us consider the groups of rational numbers of that interval, referring to the first group the numbers 0 and 1, to the second, the number $\frac{1}{2}$, to the third, all the irreducible numbers $\frac{p}{q}$

Evidently, any point of this interval is a limit point of that sequence.

3.4.4. Separation of a convergent subsequence. The results obtained in the previous subsection lead to the following *fundamental* theorem.

Theorem 3.17 (Bolzano-Weierstrass theorem)*. *A bounded sequence of numbers always contains a convergent subsequence.*

Proof. Since the sequence is bounded, it has at least one limit point x. Hence it contains a subsequence converging to point x (see Definition 2 of a limit point).

Remark 1. *Every bounded sequence always contains a monotone subsequence.* Indeed, in accordance with the Bolzano-Weierstrass theorem a bounded sequence always contains a convergent subsequence and, by virtue of the remark made in 3.4.1, this subsequence always contains a monotone subsequence.

Remark 2. *Let $\{x_n\}$ be a bounded sequence whose terms are within the interval $[a, b]$. Then the limit c of any convergent subsequence $\{x_{k_n}\}$ is also in the interval $[a, b]$.* Indeed, since $a \leqslant x_{k_n} \leqslant b$, it follows from Corollary 2 of Theorem 3.13 that the inequalities $a \leqslant c \leqslant b$ are satisfied. And this means that c lies in the interval $[a, b]$.

Take note of the fact that in particular cases an unbounded sequence can also contain a convergent subsequence. For example, the sequence $1, \frac{1}{2}, 2, \frac{1}{3}, \ldots, n, \frac{1}{n+1}, \ldots$ is unbounded whereas the subsequence $\frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}, \ldots$ of its terms with even indices converges. But not every unbounded sequence possesses a convergent subsequence. For instance, any subsequence of the unbounded sequence $1, 2, \ldots, n, \ldots$ diverges. Therefore, the Bolzano-Weierstrass theorem cannot be generalized to all unbounded sequences.

The following proposition is an analogue of this theorem for unbounded sequences.

Lemma 4. *Every unbounded sequence contains an infinitely large subsequence.*

Proof. Suppose $\{x_n\}$ is an unbounded sequence. Then, there is a term x_{k_1} of that sequence satisfying the condition $|x_{k_1}| > 1$, a term x_{k_2} of that sequence satisfying the condition $|x_{k_2}| > 2$, $k_2 > k_1$, ..., a term x_{k_n} of the sequence satisfying the condition $|x_{k_n}| > n$, $k_n > k_{n-1}$, etc. It is evident that the subsequence $x_{k_1}, x_{k_2}, \ldots, x_{k_n}, \ldots$ is infinitely large.

---

with denominator 3, and, generally, to group n we shall put all the irreducible fractions from the interval $[0, 1]$ with denominator n. Every rational number will evidently appear in one of the groups, and each group will have only a finite number of rational numbers. Let us write out in succession the terms of the first group, the terms of the second group, then those of the third group, etc. The result will be the desired subsequence.

* Bernhard Bolzano (1781-1848), a Czech philosopher and mathematician, Karl Weierstrass (1815-1897), a German mathematician.

Lemma 4 and the Bolzano-Weierstrass theorem yield the following assertion.

Lemma 5. Quite an arbitrary sequence contains either a convergent or an infinitely large subsequence.

Remark 3. The results obtained in the present subsection make it possible to extent somewhat the concepts of a limit point and of the limit superior and the limit inferior of a sequence.

We shall say that $+\infty$ ($-\infty$) is the limit point of the sequence $\{x_n\}$ if that sequence contains an infinitely large subsequence consisting of positive (negative) terms.

Within the framework of this extended interpretation of a limit point, sequences may have, besides the finite limit points, two more limit points $+\infty$ and $-\infty$. In this case Lemma 5 allows us to assert that *a completely arbitrary sequence contains at least one limit point**.

Assuming, naturally, that $+\infty$ and $-\infty$ are connected with any finite real number x by the relation $-\infty < x < +\infty$, we shall verify the fact that *a completely arbitrary sequence possesses both the limit superior and the limit inferior* (i.e. there are both the largest and the least limit points).

For definiteness, let us establish the existence of the limit superior.

By virtue of Remark 1 to Theorem 3.16, it is sufficient to consider the case where the sequence $\{x_n\}$ *is not bounded.* If $\{x_n\}$ is not bounded from above either, we can separate from it an infinitely large subsequence whose all terms are positive and, consequently, $+\infty$ is a limit point and, therefore, the limit superior of $\{x_n\}$.

It remains to consider the case where the unbounded sequence $\{x_n\}$ is bounded from above, that is, where there is a real number M such that all the terms of $\{x_n\}$ satisfy the condition $x_n \leqslant M$. Since in the case being discussed the sequence $\{x_n\}$ is not bounded from below, we can separate from it an infinitely large sequence whose all terms are negative and this means that $-\infty$ is a limit point of the sequence under consideration.

If, moreover, the sequence being considered does not possess a single finite limit point, then $-\infty$ is the only limit point and, consequently, is the limit superior of the sequence under consideration.

It remains to consider the case where the sequence being discussed possesses, besides $-\infty$, at least one more finite limit point x_0 and to prove that in this case as well this sequence possesses the limit superior. Since all the terms of $\{x_n\}$ satisfy the condition $x_n \leqslant M$, by virtue of Theorem 3.13, x_0, too, satisfies the condition $x_0 \leqslant M$. We fix an arbitrary $\varepsilon > 0$. Since there are infinitely many terms of the sequence $\{x_n\}$ in the ε-neighbourhood of x_0, there are infinitely many terms of the sequence $\{x_n\}$ on the closed interval $[x_0 - \varepsilon, M]$ as well.

Let us isolate, from the sequence $\{x_n\}$ a subsequence of the terms lying on the closed interval $[x_0 - \varepsilon, M]$. The isolated subsequence is bounded. Therefore, by virtue of Remark 1 to Theorem 3.16, it possesses the limit superior, i.e. the largest limit point $\bar{x}$. It is evident that $\bar{x} \geqslant x_0$ and is a limit point of the whole sequence $\{x_n\}$, too. It is also evident that the sequence $\{x_n\}$ does not have limit points exceeding $\bar{x}$, for if some number $\bar{\bar{x}}$ exceeding $\bar{x}$ were a limit point of the sequence $\{x_n\}$, then, all the terms of the sequence $\{x_n\}$ exceeding the number $x_0 - \varepsilon$, being the terms of the isolated subsequence, the number $\bar{\bar{x}}$ would be a limit point of the isolated subsequence as well and that subsequence would not have limit points exceeding $\bar{x}$.

Thus we infer that the number $\bar{x}$ is the largest limit point of the sequence under consideration.

We have proved the existence of the limit superior of a completely arbitrary sequence.

The proof of the existence of the limit inferior is quite similar.

* Either finite or infinite.

3.4.5. A necessary and sufficient condition of convergence of a sequence. When we use the definition of the convergence of a sequence to decide whether the sequence $\{x_n\}$ is convergent or not, we have to estimate the difference between the terms x_n of the sequence and its supposed limit a. In other words, we have to foresee the value of the limit a of the sequence.

It is natural to indicate the "interior" test of convergence of a sequence allowing the question as to its convergence to be decided only on the basis of the magnitude of its terms. This test will be established in the present subsection. To formulate the test, we shall introduce the notion of a fundamental sequence.

Definition. *The sequence $\{x_n\}$ is referred to as* **fundamental** *if for any positive ε there is a number N such that for all the numbers n satisfying the condition $n \geqslant N$ and for all the natural numbers p ($p = 1, 2, \ldots$) the inequality*

$$|x_{n+p} - x_n| < \varepsilon$$

holds true.

The main problem discussed in this subsection is the proof of the test for convergence of a sequence (the so-called Cauchy* test for convergence): *for a sequence to be convergent it is necessary and sufficient that it should be fundamental.*

Before proving Cauchy's test we shall prove some auxiliary propositions which are themselves of interest.

Theorem 3.18. *For the sequence $\{x_n\}$ to be convergent, it is necessary and sufficient that it should be bounded and that its limit superior $\overline{x}$ and limit inferior $\underline{x}$ should coincide.*

Proof. (1) **Necessity.** Suppose the sequence $\{x_n\}$ converges. Then, by virtue of Theorem 3.8, it is *bounded* and has a single limit point (in accordance with Lemma 3 in 3.4.2). Thus $\underline{x} = \overline{x}$.

(2) **Sufficiency.** Corollary 2 of Theorem 3.16 states that for any positive ε the interval $(\underline{x} - \varepsilon, \overline{x} + \varepsilon)$ contains all the terms of the sequence $\{x_n\}$ from a certain number onward. Since $\underline{x} = \overline{x} = x$, the indicated interval coincides with the ε-neighbourhood of point x, that is the number x is a limit of the sequence $\{x_n\}$ (see Remark 1 to 3.2.1).

Let us now establish an important property of a fundamental sequence following directly from its definition.

*For any positive number ε there is a term x_N of a fundamental sequence whose ε-neighbourhood contains all the terms of the sequence beginning with number N. In other words, outside the interval $(x_N - \varepsilon, x_N + \varepsilon)$ there are not more than a finite number of the terms of the sequence***.

* Augustin Louis Cauchy (1789-1857), a French mathematician.

** Note that the indicated property is equivalent to the definition of a fundamental sequence.

Indeed, it follows from the definition of a fundamental sequence that for any positive ε there is a number N such that for all natural p ($p = 1, 2, 3, \ldots$) the inequality $|x_{N+p} - x_N| < \varepsilon$ holds true, this inequality implying that the ε-neighbourhood of the term x_N contains all the terms of the sequence beginning with the number N.

This property enables us to assert that the fundamental sequence is bounded. In fact, assume that ε is some fixed positive number and x_N is a term whose ε-neighbourhood contains all the terms of the sequence beginning with the number N. Then only the terms $x_1, x_2, \ldots, x_{N-1}$ can be outside the ε-neighbourhood. Let us put $A = \max \{|x_1|, |x_2|, \ldots, |x_{N-1}|, |x_N - \varepsilon|, |x_N + \varepsilon|\}$.* Then the closed interval $[-A, +A]$ contains the numbers $x_1, x_2, \ldots, x_{N-1}, x_N - \varepsilon, x_N + \varepsilon$, and, consequently, all the points of the ε-neighbourhood of the term x_N. It follows that all the terms of the fundamental sequence are within the interval $[-A, +A]$, which means that the sequence is bounded.

Now we pass to the proof of the basic assertion discussed in this subsection.

Theorem 3.19. (Cauchy's test for convergence of a sequence. *For the sequence $\{x_n\}$ to be convergent it is necessary and sufficient that it should be fundamental.*

Proof. (1) **Necessity.** Suppose the sequence $\{x_n\}$ converges and x is its limit value. It is required to prove that the sequence is fundamental. Let us take *any* positive number ε. It follows from the definition of a convergent sequence that for the positive number $\frac{\varepsilon}{2}$ there can be found a number N such that for $n \geqslant N$ there holds an inequality

$$|x_n - x| < \frac{\varepsilon}{2}.$$

If p is any natural number, then for $n \geqslant N$ the inequality

$$|x_{n+p} - x| < \frac{\varepsilon}{2}$$

also holds true. Since the absolute value of the sum of two quantities is not greater than the sum of their absolute values, it follows from the last two inequalities that for $n \geqslant N$ and for all the natural numbers p

$$|x_{n+p} - x_n| = |(x_{n+p} - x) + (x - x_n)| \leqslant |x_{n+p} - x| + |x_n - x| < \varepsilon.$$

Thus we have established that the sequence $\{x_n\}$ is fundamental.

* In terms of geometry this means that A is equal to the maximum of the distances from the reference point 0 to the points $x_1, x_2, \ldots, x_{N-1}, x_N - \varepsilon, x_N + \varepsilon$.

(2) **Sufficiency.** Suppose $\{x_n\}$ is a fundamental sequence. It is required to prove that the sequence converges. In accordance with Theorem 3.18, for that purpose it is sufficient to prove the boundedness of the sequence $\{x_n\}$ and the equality of its limit superior $\bar{x}$ and limit inferior $\underline{x}$. The boundedness of the fundamental sequence has been established above.

To prove the equality of the limit superior $\bar{x}$ and the limit inferior $\underline{x}$, we shall make use of the following property of a fundamental sequence established earlier: for any positive number ε there can be found a term x_N such that outside the interval $(x_N - \varepsilon, x_N + \varepsilon)$ there are not more than a finite number of the terms of the sequence. We know from Corollary 1 of Theorem 3.16 that the interval $(x_N - \varepsilon, x_N + \varepsilon)$ contains the interval $(\underline{x}, \bar{x})$ and therefore $\bar{x} - \underline{x} \leqslant 2\varepsilon$, whence $\underline{x} = \bar{x}$ since ε is arbitrary. Thus we have established that the sequence is convergent, and, hence, have proved the theorem.

Example. Let us use Cauchy's test to establish the convergence of the following $\{x_n\}$ sequence:

$$x_n = a_1 + a_2 + \ldots + a_n,$$

where a_k $(k = 1, 2, 3, \ldots)$ are arbitrary real numbers satisfying the condition $|a_k| \leqslant q^k$ and q is a number lying in the interval $0 < q < 1$.

Let n be any number and p, any natural number. Then, evidently,

$$|x_{n+p} - x_n| = |a_{n+1} + a_{n+2} + \ldots + a_{n+p}| \leqslant |a_{n+1}| +$$
$$+ |a_{n+2}| + \ldots + |a_{n+p}| \leqslant q^{n+1} + q^{n+2} + \ldots + q^{n+p} =$$
$$= \frac{q^{n+1} - q^{n+1+p}}{1-q} < \frac{q^{n+1}}{1-q}.$$

Taking into account that the sequence $\{q^n\}$ is infinitesimal (see Example 1 from 3.1.3), we can assert that *for any* positive ε there is a number N such that

$$q^{n+1} < \varepsilon (1 - q) \qquad (\text{for } n \geqslant N).$$

Hence, for $n \geqslant N$ and for any natural p

$$|x_{n+p} - x_n| < \frac{q^{n+1}}{1-q} < \varepsilon,$$

i.e. the sequence $\{x_n\}$ is *fundamental* and converges in accordance with Theorem 3.19.

3.4.6. Some properties of arbitrary number sets. In this subsection we shall discuss some properties of *arbitrary number sets*. A part of these properties are similar to those of number sequence.

The concept of a *set bounded above (below)* was introduced in 2.1.5. Let us now agree to call the set $\{x\}$ *bounded* if the set has both an upper bound and a

lower bound, that is, if there are two real numbers m and M such that each member x of the set $\{x\}$ satisfies the inequality $m \leqslant x \leqslant M$. The set $\{x\}$ is to be called *finite* or *infinite* depending on whether the number of the members constituting the set is finite or infinite.

The point x of an infinite straight line is called a limit point of the set $\{x\}$ if any ε-neighbourhood of point x contains an infinite number of members of that set.

The point $\bar{x}$ (point $\underline{x}$) will be referred to as the *upper (lower) limit point* of the set $\{x\}$ if that point is the limit point of the set $\{x\}$ and not a single point larger than $\bar{x}$ (smaller than $\underline{x}$) is a limit point of the set.

Repeating almost word for word the proof of Theorem 3.16 with the only substitution of the term "set $\{x\}$" for the term "sequence $\{x_n\}$", we shall arrive at the following assertion: *every bounded infinite set has at least one limit point.*

Repeating verbatim the argument used in Remark 1 to Theorem 3.16, we shall see that *every bounded infinite set has both an upper limit point and a lower limit point.* And here is a sequence of the assertions stated above: *convergent sequence can be isolated from the members of every bounded infinite set.*

There is also a concept of a *subset* used alongside with the concept of a set. The set $\{x'\}$ is called a *subset of the set* $\{x\}$ if the set $\{x\}$ contains all the members of $\{x'\}$. For instance, the set of all even integers is a subset of the set of all integers.

Two sets $\{x\}$ and $\{y\}$ are *equivalent* if a one-to-one correspondence* can be established between the members of those sets. Note that two finite sets are equivalent if and only if the sets contain the same number of members. Here is an example of two equivalent infinite sets. It is easy to see that the set $\{x\}$ whose members are the positive even members $2, 4, 6, \ldots, 2n, \ldots$ is equivalent to the set $\{y\}$ whose members are the natural numbers $1, 2, 3, \ldots, n, \ldots$. Indeed, we can establish a one-to-one correspondence between the members of these sets associating the member $2n$ of the set $\{x\}$ with the member n of the set $\{y\}$. Pay attention to the fact that the set $\{x\}$ being considered is a subset of the set $\{y\}$. Thus the infinite set $\{y\}$ turns out to be equivalent to its subset $\{x\}$.**

From the variety of sets we shall separate the following two important types:

1°. *Every set equivalent to the set of all natural numbers* $1, 2, 3, \ldots, n, \ldots$ *will be called countable.* It follows from the definition of a countable set that all the members of that set can be enumerated.

2°. *Every set equivalent to the set of all real numbers in the interval* $(0, 1)$ *will be referred to as the set of the power of continuum.*

Here are some examples of countable sets and sets of the power of continuum. The set of positive even numbers $2, 4, 6, \ldots, 2n \ldots$ discussed above is an example of a countable set. The set of all rational numbers in the interval $[0, 1]$ can serve as another example of a countable set for, as was proved in the second note on p. 79, that set can be arranged as a sequence without recurrence, i.e. it can be enumerated. The set of all real numbers (an infinite interval) is an example of a set of the power of continuum. Indeed, the function $y = \cot \pi x$*** establishes a one-to-one correspondence between the points of the interval $[0 < x < 1]$ and the points of an infinite interval.

---

* A one-to-one correspondence between the members of two sets is such a correspondence which associates every member of one set with only one member of the other set so that each member of the second set corresponds to only one member of the first set.

** It is easy to see that any infinite set is equivalent to its certain subset not coinciding with the entire set. This fact may be accepted as the definition of an infinite set.

*** The reader is sure to be familiar with the function $y = \cot \pi x$ from the elementary course. A rigorous construction of trigonometric functions is discused in Chap. 4.

In conclusion we shall prove that a *set of the power of continuum is not equivalent to a countable set.* To do that it is sufficient to prove that the set of all real numbers in the interval (0, 1) cannot be enumerated. Let us proceed from the contrary, that is, let us prove that all the real numbers in the interval (0, 1) can be enumerated. Then, writing these numbers as infinite decimal fractions, we obtain a sequence

$$x_1 = 0,\ a_{11}\ a_{12} \ldots a_{1n} \ldots,$$
$$x_2 = 0,\ a_{21}\ a_{22} \ldots a_{2n} \ldots,$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$x_n = 0,\ a_{n1}\ a_{n2} \ldots a_{nn} \ldots$$
$$\ldots\ldots\ldots\ldots\ldots$$

Let us write now a real number $x = 0, b_1, b_2, \ldots b_n \ldots$, where b_1 is any digit different from a_{11}, 0 and 9, b_2 is any digit different from a_{22}, 0 and 9 and, generally, b_n is any digit different from a_{nn}, 0 and 9.

Since the number x does not contain any zeros or nines in the decimal places, it does not belong to the class of rational numbers which can be represented as infinite decimal fractions* by the two methods discussed above. But in that case the number x is sure to differ from *all* the numbers $x_1, x_2, \ldots, x_n, \ldots$, for the coincidence of the number x with any of the numbers x_n would imply the coincidence of b_n and a_{nn}.

Mathematicians were long interested in the question of whether there existed infinite sets $\{x\}$ not equivalent to either a countable set or a set of the power of continuum, but equivalent to a part of a set of the power of continuum. American mathematician P. Cohen proved in 1963 that the hypothesis concerning the existence of such a set does not depend on the rest of the axioms of the set theory. This means that it is possible to construct a consistent theory of sets by postulating either the existence of such a set or its absence (see P. Cohen's *Set Theory and the Continuum*, New York, 1966).

SUPPLEMENT 1 TO CHAPTER 3

STOLZ' THEOREM

The following statement turns out to be useful in many cases when the convergence of the quotient $\left\{\frac{x_n}{y_n}\right\}$ upon the division of the sequence $\{x_n\}$ by $\{y_n\}$ is investigated.

Stolz' theorem. *Let $\{y_n\}$ be an increasing infinitely large sequence and let the sequence $\left\{\frac{x_n - x_{n-1}}{y_n - y_{n-1}}\right\}$ converge and possess a limit a. Then the sequence $\left\{\frac{x_n}{y_n}\right\}$ converges and has a limit a. Thus*

$$\lim_{n\to\infty} \frac{x_n}{y_n} = \lim_{n\to\infty} \frac{x_n - x_{n-1}}{y_n - y_{n-1}}.$$

Proof. Since the sequence $\left\{\frac{x_n - x_{n-1}}{y_n - y_{n-1}}\right\}$ converges and has a limit a, sequence $\{\alpha_n\}$, where $\alpha_n = \frac{x_n - x_{n-1}}{y_n - y_{n-1}} - a$, is infinitesimal. Let $\overline{N}$ be any fixed number

* See 2.1.3.

and $n > \overline{N}$. Using the expression for α_n, we shall consider the following equalities:

$$x_{\overline{N}+1} - x_{\overline{N}} = a\,(y_{\overline{N}+1} - y_{\overline{N}}) + \alpha_{\overline{N}+1}\,(y_{\overline{N}+1} - y_{\overline{N}}),$$
$$x_{\overline{N}+2} - x_{\overline{N}+1} = a\,(y_{\overline{N}+2} - y_{\overline{N}+1}) + \alpha_{\overline{N}+2}\,(y_{\overline{N}+2} - y_{\overline{N}+1}),$$
$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$
$$x_{n-1} - x_{n-2} = a\,(y_{n-1} - y_{n-2}) + \alpha_{n-1}\,(y_{n-1} - y_{n-2}),$$
$$x_n - x_{n-1} = a\,(y_n - y_{n-1}) + \alpha_n\,(y_n - y_{n-1}).$$

Summing up the equalities, we obtain

$$x_n - x_{\overline{N}} = ay_n - ay_{\overline{N}} + \alpha_{\overline{N}+1}\,(y_{\overline{N}+1} - y_{\overline{N}}) + \alpha_{\overline{N}+2}\,(y_{\overline{N}+2} - y_{\overline{N}+1}) +$$
$$+ \dots + \alpha_{n-1}\,(y_{n-1} - y_{n-2}) + \alpha_n\,(y_n - y_{n-1}).$$

Inasmuch as $\{y_n\}$ is an increasing infinitely large sequence, its terms are positive beginning with a certain number. We shall assume that $y_n > 0$ when $n \geqslant \overline{N}$. Then we obtain from the last equality

$$\frac{x_n}{y_n} - a = \frac{x_{\overline{N}} - ay_{\overline{N}}}{y_n} +$$
$$+ \frac{\alpha_{\overline{N}+1}\,(y_{\overline{N}+1} - y_{\overline{N}}) + \alpha_{\overline{N}+2}\,(y_{\overline{N}+2} - y_{\overline{N}+1}) + \dots + \alpha_n\,(y_n - y_{n-1})}{y_n}$$

Since the sequence $\{y_n\}$ is increasing, the differences $y_{k+1} - y_k$, $k = \overline{N}, \overline{N} + 1, \dots, n - 1$ are positive. Therefore we have, from the last relation,

$$\left|\frac{x_n}{y_n} - a\right| \leqslant \left|\frac{x_{\overline{N}} - ay_{\overline{N}}}{y_n}\right| +$$
$$+ \frac{|\alpha_{\overline{N}+1}|\,(y_{\overline{N}+1} - y_{\overline{N}}) + |\alpha_{\overline{N}+2}|\,(y_{\overline{N}+2} - y_{\overline{N}+1}) + \dots + |\alpha_n|\,(y_n - y_{n-1})}{y_n} \tag{3.8}$$

We shall prove now that the sequence $\left\{\frac{x_n}{y_n}\right\}$ converges and has a limit a. For our purpose it is sufficient to prove that for any positive ε there is a number N such that for $n \geqslant N$ the inequality $\left|\frac{x_n}{y_n} - a\right| < \varepsilon$ is satisfied. First, proceeding from a given positive ε we shall choose a number $\overline{N}$ such that for $n \geqslant \overline{N}$ the inequality $|\alpha_n| < \frac{\varepsilon}{2}$ is complied with (this is possible since the sequence $\{\alpha_n\}$ is infinitesimal). Now we must take a number $N \geqslant \overline{N}$ such that for $n \geqslant N$ there holds an inequality $\left|\frac{x_{\overline{N}} - ay_{\overline{N}}}{y_n}\right| < \frac{\varepsilon}{2}$. Such a choice of number N is possible since the number $x_{\overline{N}} - ay_{\overline{N}}$ is fixed and the sequence $\{y_n\}$ is infinitely large and, hence, the sequence $\left\{\frac{x_{\overline{N}} - ay_{\overline{N}}}{y_n}\right\}$ is infinitesimal. Suppose now that $n \geqslant N$. Inequality (3.8) yields

$$\left|\frac{x_n}{y_n}-a\right|<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}\cdot\frac{(y_{\overline{N}+1}-y_{\overline{N}})+(y_{\overline{N}+2}-y_{\overline{N}+1})+\ldots+(y_n-y_{n-1})}{y_n},$$

or

$$\left|\frac{x_n}{y_n}-a\right|<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}\,\frac{y_n-y_{\overline{N}}}{y_n}.$$

Since $y_n-y_{\overline{N}}\leqslant y_n$ and $y_n>0$ for $n\geqslant N$, we have $\dfrac{y_n-y_{\overline{N}}}{y_n}\leqslant 1$. Therefore, it follows from the last inequality that for $n\geqslant N$

$$\left|\frac{x_n}{y_n}-a\right|<\varepsilon.$$

That was the proof of the theorem.

Remark. *If $\{y_n\}$ is an increasing infinitely large sequence and the sequence $\left\{\dfrac{x_n-x_{n-1}}{y_n-y_{n-1}}\right\}$ is also infinitely large and tends to infinity of a definite sign, then the sequence $\left\{\dfrac{x_n}{y_n}\right\}$ is infinitely large.* Indeed, assume that

$$\frac{x_n-x_{n-1}}{y_n-y_{n-1}}=A_n.$$

The sequence $\{A_n\}$ is infinitely large. We have for $n\geqslant\overline{N}$

$$\begin{aligned}
&x_{\overline{N}+1}-x_{\overline{N}}=A_{\overline{N}+1}(y_{\overline{N}+1}-y_{\overline{N}}),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\\
&x_n-x_{n-1}=A_n(y_n-y_{n-1}).
\end{aligned}$$

Summing up these equalities, we find

$$x_n-x_{\overline{N}}=A_{\overline{N}+1}(y_{\overline{N}+1}-y_{\overline{N}})+\ldots+A_n(y_n-y_{n-1}),$$

whence

$$\frac{x_n}{y_n}=\frac{A_{\overline{N}+1}(y_{\overline{N}+1}-y_{\overline{N}})+\ldots+A_n(y_n-y_{n-1})}{y_n}+\frac{x_{\overline{N}}}{y_n}.$$

This relation yields

$$\left|\frac{x_n}{y_n}\right|\geqslant\left|\frac{A_{\overline{N}+1}(y_{\overline{N}+1}-y_{\overline{N}})+\ldots+A_n(y_n-y_{n-1})}{y_n}\right|-\left|\frac{x_{\overline{N}}}{y_n}\right|. \tag{3.9}$$

We shall assume for the sake of definiteness that for $n\geqslant\overline{N}$ the terms of the sequences $\{y_n\}$ and $\{A_n\}$ are positive. Proceeding from a specified positive A we shall choose a number $\overline{N}$ such that for $n\geqslant\overline{N}$ the inequality $A_n>4A$ is satisfied and then such an $N\geqslant\overline{N}$ that for $n\geqslant N$

$$\left|\frac{x_{\overline{N}}}{y_n}\right|<A,\quad\frac{y_{\overline{N}}}{y_n}<\frac{1}{2}.$$

The choice of such an N is possible because the sequences $\{A_n\}$ and $\{y_n\}$ are infinitely large and their terms, from a certain number onward, are positive. It is

evident that when $n \geqslant N$ we have, from (3.9),

$$\left|\frac{x_n}{y_n}\right| > 4A\,\frac{(y_{\overline{N}+1}-y_{\overline{N}})+\ldots+(y_n-y_{n-1})}{y_n} - \left|\frac{x_{\overline{N}}}{y_n}\right|,$$

or

$$\left|\frac{x_n}{y_n}\right| > 4A\left(1-\frac{y_{\overline{N}}}{y_n}\right)-A > A.$$

Thus we see that the sequence $\left\{\frac{x_n}{y_n}\right\}$ is infinitely large.

Let us consider some examples.

1°. We shall prove that if the sequence $\{a_n\}$ converges and has a limit a, then the sequence $\left\{\frac{a_1+a_2+\ldots+a_n}{n}\right\}$ of the arithmetic mean values of the terms of the sequence $\{a_n\}$ converges to the same limit a.* Indeed, if we put $a_1+a_2+\ldots+a_n=x_n$ and $y_n=n$, then $\frac{x_n-x_{n-1}}{y_n-y_{n-1}}=a_n$. Since $\lim\limits_{n\to\infty}\frac{x_n-x_{n-1}}{y_n-y_{n-1}}=\lim\limits_{n\to\infty}a_n$ exists, it follows, from Stolz' theorem, that

$$\lim_{n\to\infty}\frac{a_1+a_2+\ldots+a_n}{n}=\lim_{n\to\infty}a_n=a.$$

2°. Let us consider now the sequence $\{a_n\}$, where

$$a_n=\frac{1^k+2^k+\ldots+n^k}{n^{k+1}}$$

and k is a positive integer.

Let us express $1^k+2^k+\ldots+n^k$ by x_n and n^{k+1} by y_n. Then the sequence $\{a_n\}$ will take on the form $\left\{\frac{x_n}{y_n}\right\}$. Having analysed the convergence of the sequence $\left\{\frac{x_n-x_{n-1}}{y_n-y_{n-1}}\right\}$, we see that

$$\frac{x_n-x_{n-1}}{y_n-y_{n-1}}=\frac{n^k}{n^{k+1}-(n-1)^{k+1}}=$$

$$=\frac{n^k}{(k+1)\,n^k-\frac{(k+1)\,k}{2}\,n^{k-1}+\ldots+(-1)^{k+1}}.$$

We divide the numerator and denominator of the last fraction by n^k and obtain

$$\frac{x_n-x_{n-1}}{y_n-y_{n-1}}=\frac{1}{k+1+\frac{1}{n}[\ldots]},$$

where in the denominator, in the square brackets, an expression is omitted whose limit for $n\to\infty$, is equal to $\left[-\frac{(k+1)k}{2}\right]$. We find, from the last formula,

$$\lim_{n\to\infty}\frac{x_n-x_{n-1}}{y_n-y_{n-1}}=\frac{1}{k+1}.$$

* This proposition was proved by Cauchy.

Consequently, from Stolz' theorem we have

$$\lim_{n\to\infty}\frac{1^k+2^k+\ldots+n^k}{n^{k+1}}=\frac{1}{k+1}.$$

3°. Let us, finally, consider the sequence $\left\{\frac{a^n}{n}\right\}$, $a>1$. Putting $a^n=x_n$ and $n=y_n$ and analysing the sequence $\left\{\frac{x_n-x_{n-1}}{y_n-y_{n-1}}\right\}$ we find

$$\lim_{n\to\infty}\frac{x_n-x_{n-1}}{y_n-y_{n-1}}=\lim_{n\to\infty}(a^n-a^{n-1})=\lim_{n\to\infty}a^n\left(1-\frac{1}{a}\right)=+\infty.$$

Therefore, in accordance with the remark to Stolz' theorem, we have

$$\lim_{n\to\infty}\frac{a^n}{n}=+\infty$$

SUPPLEMENT 2 TO CHAPTER 3

ON THE RATE OF CONVERGENCE OF A SEQUENCE APPROACHING $\sqrt{a}$

We have proved in 3.3 of this chapter that the limit of the sequence $\{x_n\}$ specified by the recurrence formula

$$x_{n+1}=\frac{1}{2}\left(x_n+\frac{a}{x_n}\right),\qquad n=1,\ 2,\ \ldots, \tag{3.10}$$

where $a>0$ and x_1 is any positive number, is equal to $\sqrt{a}$. We can take any term x_{n+1} of this sequence as an approximate value of $\sqrt{a}$. Naturally, in this case we should decide on the number n of iterations* ensuring the approximation of $\sqrt{a}$ with a specified degree of accuracy.

Let us now turn to the sequence $\{x_n\}$ that can be determined by the recurrence formula (3.10). We shall call the term x_n of the sequence the *nth approximation of the number* $\gamma=\sqrt{a}$. The quantity

$$\varepsilon_n=\frac{x_n-\gamma}{\gamma} \tag{3.11}$$

will be referred to as the *relative error* of the nth approximation.

The following *assertion* concerning the estimation of the relative error ε_{n+1} in terms of the relative error of the first approximation is valid.

Suppose x_1 *is chosen so that* $|\varepsilon_1|<\frac{1}{2}$. *Then for any* $n\geqslant 1$ *there hold inequalities*

$$0\leqslant\varepsilon_{n+1}\leqslant\varepsilon_1^{2^n}. \tag{3.12}$$

* *Iteration* (from the Latin iteratio meaning repetition) is a result of repetitive application of a mathematical operation. In the example considered, one iteration is the calculation of x_{n+1} from the given x_n by the recurrence formula (3.10).

Proof. It follows from formula (3.11) that

$$x_n = \gamma (1 + \varepsilon_n). \tag{3.13}$$

Considering formulas (3.10) and (3.13) and the equality $a/\gamma = \gamma$, we obtain

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right) = \frac{1}{2}\left[\gamma(1+\varepsilon_n) + \frac{a}{\gamma(1+\varepsilon_n)}\right] =$$

$$= \gamma\left[1 + \frac{\varepsilon_n^2}{2(1+\varepsilon_n)}\right].$$

Since $x_{n+1} = \gamma (1 + \varepsilon_{n+1})$, we shall evidently have

$$\varepsilon_{n+1} = \frac{1}{2(1+\varepsilon_n)}\varepsilon_n^2. \tag{3.14}$$

By the hypothesis $|\varepsilon_1| < \frac{1}{2}$. This yields inequalities $0 < \frac{1}{2(1+\varepsilon_1)} < 1$. But then it follows from (3.14) that $\varepsilon_2 \geqslant 0$ for $n = 1$. Making use of relation (3.14) for $n = 2, 3, \ldots$, we make sure that ε_{n+1} is nonnegative for $n \geqslant 1$.

The inequality $\varepsilon_{n+1} \leqslant \varepsilon_n^2$ for any $n \geqslant 1$ follows from equality (3.14), from relations $0 < \frac{1}{2(1+\varepsilon_1)} < 1$ and from the fact that ε_n is nonnegative for any $n > 1$. From this inequality immediately follows the right-hand inequality (3.12). We have thus proved the assertion.

Studying inequalities (3.12) we see that the relative error ε_{n+1} in the calculation of $\sqrt{a}$ after n iterations is estimated in terms of the relative error ε_1 of the first approximation x_1 and the number n of iterations. We shall see later on that with $a > 1$* the first approximation x_1 can be chosen so that the absolute value of ε_1 will not exceed 0.05. It is evident that with such a choice of x_1 the relative error ε_1 will satisfy the conditions of the assertion just proved. It is also clear that this is the solution of the question concerning the number n of iterations that ensures the approximation of $\sqrt{a}$ with the specified relative error ε: *that number n can be found from the formula***

$$(0.05)^{2^n} < \varepsilon. \tag{3.15}$$

Thus we suppose that $a > 1$. Let us represent the number a in the following form

$$a = 2^{2k+l}M, \tag{3.16}$$

where k is a nonnegative integer, the number l is either zero or unity, and the number M complies with the conditions

$$1 \leqslant M < 2. \tag{3.17}$$

Note that the representation of the number a in form (3.16) is unique.

Let us choose x_1 as follows

$$x_1 = 2^k\left(\frac{1}{3}2^l M + \frac{17}{24}\right). \tag{3.18}$$

We shall now ascertain that *for any M* satisfying conditions (3.17) the first approximation x_1 calculated by formula (3.18) yields a relative error ε_1 in computing $\gamma = \sqrt{a}$ not exceeding 0.05 in its absolute value.

* If $a < 1$, then $a = \frac{1}{b}$, where $b > 1$ and $\sqrt{a}$ is equal to $\frac{1}{\sqrt{b}}$.

** The validity of the formula follows directly from relation (3.12).

For our purpose let us turn to a strict expression of the relative error $\varepsilon_1 = \frac{x_1 - \gamma}{\gamma}$. Since $\gamma = 2^k \sqrt{2^i M}$ in accordance with (3.16), it follows from the expression for ε_1 and formula (3.18) that

$$\varepsilon_1 = \frac{\frac{1}{3} 2^i M + \frac{17}{24} - \sqrt{2^i M}}{\sqrt{2^i M}}. \tag{3.19}$$

Since the number i is equal either to zero or to unity and $M \geqslant 1$, we have $\sqrt{2^i M} \geqslant 1$. This expression and (3.19) yield inequality

$$|\varepsilon_1| \leqslant \left| \frac{1}{3} 2^i M + \frac{17}{24} - \sqrt{2^i M} \right|. \tag{3.20}$$

Let us designate $\sqrt{2^i M}$ by X. Since $1 \leqslant M < 2$ and the number i is either zero or unity, all the admissible values of X are sure to lie within the closed interval [1, 2]:

$$1 \leqslant X \leqslant 2. \tag{3.21}$$

Using the symbol X we have introduced for $\sqrt{2^i M}$, we shall rewrite inequality (3.20) in the following form:

$$|\varepsilon_1| \leqslant \left| \frac{1}{3} X^2 - X + \frac{17}{24} \right|. \tag{3.22}$$

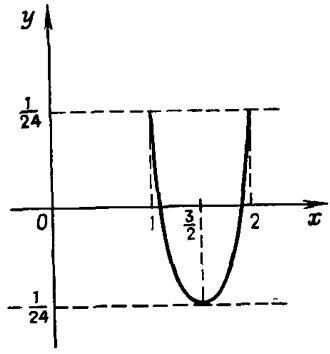


Fig. 3.4

By virtue of (3.22) the maximum value $|\varepsilon_1|$ does not exceed the maximum value of $\left| \frac{1}{3} X^2 - X + \frac{17}{24} \right|$ for the values of X satisfying conditions (3.21). To elucidate the question concerning this maximum value, let us consider the graph of the function $f(X) = \frac{1}{3} X^2 - X + \frac{17}{24}$. It is known from the course of elementary mathematics that the graph of this function is a parabola whose vertex is the point $X = \frac{3}{2}$ (Fig. 3.4)*. Since $f(1) = f(2) = \frac{1}{24}$ and $f\left(\frac{3}{2}\right) = -\frac{1}{24}$, it is clear that for the values of X satisfying conditions (3.21) the values of $f(\ )$ lie between $-\frac{1}{24}$ and $\frac{1}{24}$. In other words,

$$|f(X)| = \left| \frac{1}{3} X^2 - X + \frac{17}{24} \right| \leqslant \frac{1}{24}.$$

The last inequality and inequality (3.22) yield the desired inequality for ε_1:

$$|\varepsilon_1| \leqslant \frac{1}{24} < 0.05.$$

Remark. Note that if the specified relative error ε is equal to 10^{-10} then to calculate the square root of *any number* $a > 1$ with such a degree of accuracy, after choosing x_1 by formula (3.18), only *three iterations* ($n = 3$) will be required because $(0.05)^{2^3} < 10^{-10}$.

* In Fig. 3.4 the scale along the y-axis is 20 times that along the x-axis.

CHAPTER 4

THE CONCEPT OF A FUNCTION. THE LIMITING VALUE OF A FUNCTION. CONTINUITY

We begin the chapter with elucidating one of the most important concepts of mathematical analysis, the concept of a function. Proceeding from the concept of the limit of a number sequence, we shall introduce a new form of a limiting process based on the concept of the limiting value (or limit) of a function. In this chapter we shall also deal with continuity of a function, a concept significant for mathematical analysis.

A considerable space is given to elucidating the property of continuity and other important properties of the basic elementary functions.

The problem of finding approximate values of elementary functions is treated in Supplement to Chapter 8.

4.1. THE CONCEPT OF A FUNCTION

4.1.1. A variable quantity and a function. We have mentioned in Chapter 1 that any real physical process is associated with at least two variable quantities whose variations are interdependent.

While considering real physical variables we come to the conclusion that it is not in every case that these quantities can take on arbitrary values. For example, the body temperature cannot be lower than $-273°C$, the velocity of a particle cannot exceed the value $3 \cdot 10^8$ ms^{-1} (i.e. the velocity of light in free space), the displacement y of a particle oscillating harmonically according to the law $y = A \sin(\omega t + \delta)$ can vary only within the closed interval $[-A, +A]$.

Mathematics abstracts from concrete physical properties of variable quantities observed in nature and considers an abstract variable* characterized only by numerical values which it can assume.

The set $\{x\}$ of all values that the given variable may take on is called the *range of variation* of that variable. The variable quantity

* It should be pointed out that the concept of a quantity is one of the basic mathematical concepts (see footnote on p. 18).

is said to be defined if its range of variation is given. In what follows we shall denote variable quantities by lower-case letters x,

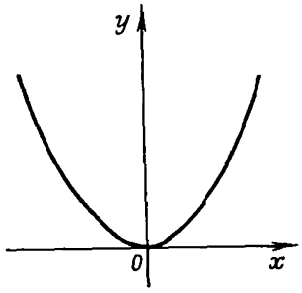


Fig. 4.1

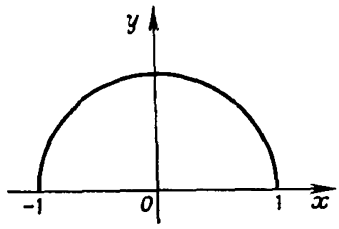


Fig. 4.2

y, u, . . ., and their ranges of variation by the symbols $\{x\}$, $\{y\}$, $\{u\}$, ...

Suppose a variable x is given whose range of variation is the set $\{x\}$.

When every value of the variable x of the set $\{x\}$ is associated with a certain number y, in accordance with a known law, then it is said that the function $y = y(x)$ or $y = f(x)$ is defined on the set $\{x\}$.

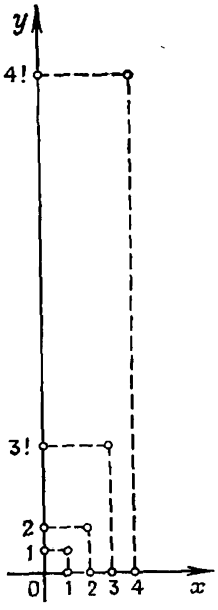


Fig. 4.3

In that case the variable x is called the *argument* and the set $\{x\}$ is referred to as the *domain of definition* of the function $y = f(x)$.

The number y corresponding to a given value of the argument x is called a *particular value of the function* at point x. The totality of all particular values of a function constitutes a definite set $\{y\}$ called the *range of the function.*

In the notation $y = f(x)$ the letter f is called the *characteristic* of the function. Various letters may be used to designate the argument, the function and its characteristic.

Here are some examples of functions.

1°. $y = x^2$. This function is defined on an infinite interval $-\infty < x < +\infty$. The range of this function is the half-line $0 \leqslant y < +\infty$ (Fig. 4.1).

2°. $y = \sqrt{1-x^2}$. The function is defined on the interval $-1 \leqslant x \leqslant +1$. The range of the function is the interval $0 \leqslant y \leqslant 1$ (Fig. 4.2).

3°. $y = n!$. This function is defined on the set of natural numbers $n = 1, 2, \ldots$. The range of this function is the set of natural numbers of the type $n!$ (Fig. 4.3).

4°. The Dirichlet* function

$$y=\begin{cases}0 & \text{if } x \text{ is an irrational number,}\\ 1 & \text{if } x \text{ is a rational number.}\end{cases}$$

This function is defined on the infinite interval $-\infty < x < +\infty$ and its range consists of two points, 0 and 1.

5°.

$$y=\operatorname{sgn} x=\begin{cases}+1 & \text{if } x>0,\\ 0 & \text{if } x=0,\\ -1 & \text{if } x<0.\end{cases}$$

(The term signem is from the Latin signum meaning sign.) This function is defined on the entire infinite interval $-\infty < x < +\infty$, and its range consists of three points, -1, 0 and $+1$ (Fig. 4.4).

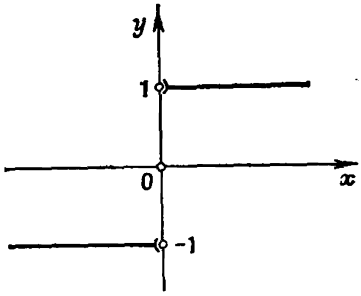


Fig. 4.4

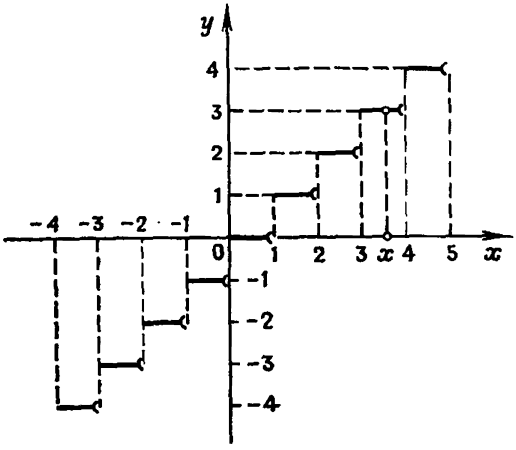


Fig. 4.5

6°. $y = [x]$, where $[x]$ denotes the integral part of the real number x. This function is defined for all the real values of x, and its range consists of integral numbers (Fig. 4.5).

4.1.2. Methods of representing a function. Here we shall dwell on some methods representing a function.

Very often the law establishing the relationship between the argument and the function is specified by means of a formula. This method of representation of a function is called *analytical*. It should be noted that a function can be specified by different formulas in various intervals of its domains of definition.

For instance, the function

$$y=\begin{cases}\sin x & \text{for } x\leqslant 0,\\ x^2 & \text{for } x>0\end{cases}$$

* Peter Gustav Lejeune Dirichlet (1805-1859), a German mathematician.

is represented analytically on the entire infinite interval (Fig. 4.6).

A *tabular method* of representing a function is very widespread. It consists in tabulating certain values of the argument and the corresponding values of the function. This method makes it possible to calculate approximately the values of the function not contained in the table which correspond to the intermediate values of the argument. To do that, a *method of interpolation* is employed consisting in the replacement of the function lying between

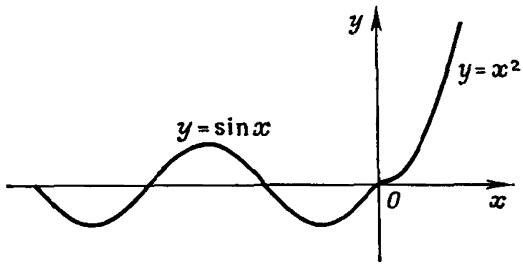


Fig. 4.6

its tabular values by some simple function (say, linear or quadratic). A railway schedule can serve as an example of a function representation. It specifies the location of a train at certain time moments. The interpolation makes possible an approximate determination of the location of the train at any intermediate moment of time.

One more method of representing a function, a *graphical method*, is used in practical measurements. By this method the relationship between the argument and the function is given by a graph (recorded by an oscillograph, for instance).

4.2. THE CONCEPT OF THE LIMITING VALUE OF A FUNCTION

4.2.1. Definition of the limiting value of a function. Let us consider a function $y = f(x)$ defined on a certain set $\{x\}$, and a point a not necessarily belonging to the set $\{x\}$ but possessing the property such that in any ε-neighbourhood of point a there are points of the set $\{x\}$ distinct from a. For instance, point a may be the end point of the interval on which the function is defined.

Definition 1. *The number b is called the* **limiting value of the function $y = f(x)$ at point $x = a$** *(or the* **limit of a function for $x \to a$***) if for any sequence $x_1, x_2, \ldots, x_n, \ldots$ of the values of the argument x, converging to a, whose terms x_n are different from a* ($x_n \neq a$), the*

* This requirement can be explained, in particular, by the fact that the function $f(x)$ may not be specified at point a.

corresponding sequence $f(x_1)$, $f(x_2)$, . . ., $f(x_n)$, . . . *of the values of the function converges to* b.

The limiting value of a function is symbolized as follows:

$$\lim_{x \to a} f(x) = b.$$

Note that the *function* $y = f(x)$ *can have only one limiting value at point* a. This follows from the fact that the sequence $\{f(x_n)\}$ can have only one limit.

Let us consider some examples.

1°. The function $f(x) = c$ possesses a limiting value at every point a of an infinite interval. Indeed, if $x_1, x_2, \ldots, x_n, \ldots$ is any sequence of the values of the argument converging to a, then the corresponding sequence of the values of the function has the form $c, c, \ldots, c, \ldots$ and, therefore, converges to c. Thus, the limiting value of that function at any point $x = a$ is equal to c.

2°. The limiting value of the function $f(x) = x$ at any point a of an infinite interval is equal to a. Indeed, in that case the sequences of the values of the argument and the function are identical and, therefore, if the sequence $\{x_n\}$ converges to a, the sequence $\{f(x_n)\}$ converges to a as well.

3°. The Dirichlet function whose values at rational points are equal to unity and at irrational points to zero does not have a limiting value at any point a of an infinite interval. Indeed, for a sequence of rational values of the argument, converging to a, the limit of the corresponding sequence of values of the function is equal to unity, and for the sequence of irrational values of the argument converging to a the limit of the corresponding sequence of values of the function is equal to zero.

In what follows we shall use one-sided limiting values of a function which are defined as follows.

We shall assume that the set $\{x\}$ on which the function $f(x)$ is defined, possesses, for any $\varepsilon > 0$, at least one member lying on the interval $(a, a + \varepsilon)$ (on the interval $(a - \varepsilon, a)$ respectively).

Definition 2. *The number* b *is referred to as the* **limit on the right (on the left)** *of the function* $f(x)$ *at point* $x = a$ *if for any sequence* $x_1, x_2, \ldots, x_n, \ldots$ *of the values of the argument* x, *converging to* a, *whose terms* x_n *are greater (smaller) than* a, *the corresponding sequence* $f(x_1), f(x_2), \ldots, f(x_n), \ldots$ *of the values of the function converges to* b.

The limit on the right of the function is denoted as

$$\lim_{x \to a+0} f(x) = b \quad \text{or} \quad f(a+0) = b.$$

The limit on the left is designated as

$$\lim_{x \to a-0} f(x) = b \quad \text{or} \quad f(a-0) = b.$$

Let us consider the function $f(x) = \operatorname{sgn} x$* as an example. The limiting values of this function on the right and on the left are zero, and $\operatorname{sgn}(0+0) = 1$, $\operatorname{sgn}(0-0) = -1$. Indeed, if $\{x_n\}$ is any sequence of the values of the argument of this function, converging to zero, whose x_n terms are greater than zero ($x_n > 0$), then $\operatorname{sgn} x_n = 1$ and consequently $\lim\limits_{n\to\infty} \operatorname{sgn} x_n = 1$. Thus we have established the validity of the equality $\operatorname{sgn}(0+0) = 1$. The proof of the equality $\operatorname{sgn}(0-0) = -1$ is quite analogous.

Remark. *If at point a the limits of the function $f(x)$ on the right and on the left are equal, then there is a limiting value of that function at point a equal to the indicated one-sided limiting values.* We shall now prove this obvious fact.

Let $\{x_n\}$ be any sequence of values of the argument of the function $f(x)$, converging to a, whose terms are not equal to a. And let $\{x'_{k_m}\}$ be a subsequence of that sequence consisting of all the terms of the sequence $\{x_n\}$ which are greater than a, and $\{x_{l_m}\}$ be a subsequence containing all the terms of $\{x_n\}$ which are smaller than a.** Since, by virtue of 3.4.1, the subsequences $\{x_{k_m}\}$ and $\{x_{l_m}\}$ converge to a, it follows from the existence of the limiting values of the function $f(x)$ on the right and on the left at point a that the sequences $\{f(x_{k_m})\}$ and $\{f(x_{l_m})\}$ have limits which are equal by the hypothesis. Let b be the limit of those sequences. For any $\varepsilon > 0$ there is a number N such that all the terms of the sequences $\{f(x_{k_m})\}$ and $\{f(x_{l_m})\}$, for which $k_m \geqslant N$ and $l_m \geqslant N$, satisfy the inequalities $|f(x_{k_m}) - b| < \varepsilon$ and $|f(x_{l_m}) - b| < \varepsilon$. Consequently, for $n \geqslant N$, the inequality $|f(x_n) - b| < \varepsilon$ is fulfilled, which means that the sequence $\{f(x_n)\}$ converges to b. Thus we have proved that the limiting value of the function $f(x)$ at point a exists and is equal to b.

We shall now formulate the definitions of the limiting values of the function in the cases where its argument x tends to infinity and to the infinity of a certain sign.

We shall assume that the set $\{x\}$ on which the function $f(x)$ is defined possesses, for any $A > 0$, at least one member lying outside of the closed interval $[-A, +A]$.

Definition 3. *The number b is referred to as the* **limiting value of the function $f(x)$ for $x \to \infty$** *(or the* **limit of the function for $x \to \infty$***) if for any infinitely large sequence of the values of its argument the corresponding sequence of the values of the function converges to b.*

The following notation is used to designate the limiting value of

---

* The definition of the function $y = \operatorname{sgn} x$ was given in 4.1.1.

** We exclude from our discussion the case where *only a finite number of terms* of the sequence $\{x_n\}$ are to the right (to the left) of point a. In that case the convergence of $\{f(x_n)\}$ is obvious.

the function for $x \to \infty$:

$$\lim_{x \to \infty} f(x) = b.$$

We shall finally assume that the set $\{x\}$ on which the function $f(x)$ is defined possesses, for any $A > 0$, at least one member x satisfying the condition $x > A$ $(x < -A)$.

Definition 4. *The number b is referred to as the limiting value of the function $f(x)$ with the argument x tending to the positive (negative) infinity, if for any infinitely large sequence of values of the argument whose terms, beginning with a certain number, are positive (negative), the corresponding sequence of the values of the function converges to b.*

The notation is

$$\lim_{x \to +\infty} f(x) = b \quad \left(\lim_{x \to -\infty} f(x) = b\right).$$

Let us consider a function $f(x) = \frac{1}{x}$ as an example. This function has a limit equal to zero for $x \to \infty$. Indeed, if $x_1, x_2, \ldots, x_n, \ldots$ is an infinitely large sequence of values of the argument, then the sequence $\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_n}, \ldots$ is infinitesimal and, therefore, has a limit equal to zero.

4.2.2. Arithmetic operations on functions possessing limiting values. Let us make sure that arithmetic operations on a function possessing a limit at point a lead to a function which also has a limit at that point. The following *fundamental* theorem is valid.

Theorem 4.1. *Let the functions $f(x)$ and $g(x)$ defined on the same set possess the limiting values b and c at point a. Then the functions $f(x) + g(x)$, $f(x) - g(x)$, $f(x) \cdot g(x)$ and $\frac{f(x)}{g(x)}$ possess, at point a, the limiting values (a quotient, under the condition $c \neq 0$) equal, respectively, to $b + c$, $b - c$, $b \cdot c$ and $\frac{b}{c}$.*

Proof. Let $x_1, x_2, \ldots, x_n, \ldots$ $(x_n \neq a)$ be an arbitrary sequence of values of the argument of the functions $f(x)$ and $g(x)$ converging to a. The corresponding sequences $f(x_1), f(x_2), \ldots, f(x_n) \ldots$ and $g(x_1), g(x_2), \ldots, g(x_n), \ldots$ of the values of these functions have the limits b and c. But then, in accordance with Theorems 3.9-3.12, the sequences $\{f(x_n)+g(x_n)\}$, $\{f(x_n)-g(x_n)\}$, $\{f(x_n)\cdot g(x_n)\}$ and $\left\{\frac{f(x_n)}{g(x_n)}\right\}$ have limits equal, respectively, to $b+c$, $b-c$, $b \cdot c$ and $\frac{b}{c}$. Since the sequence $\{x_n\}$ is arbitrary, this means that $\lim_{x \to a} [f(x)+g(x)] = b+c$, $\lim_{x \to a} [f(x)-g(x)] = b-c$, $\lim_{x \to a} [f(x) \cdot g(x)] = b \cdot c$, $\lim_{x \to a} \frac{f(x)}{g(x)} = \frac{b}{c}$. The theorem has been proved.

Let us use this theorem to determine the limiting values of poly-

nomials and irreducible algebraic fractions*. We can make the following statement.

At every point a of an infinite interval there exist limiting values of polynomials and irreducible algebraic fractions which are equal to the particular values of those functions at the indicated point (in the case of an algebraic fraction, a must not be a root of the denominator).

Indeed, by virtue of Theorem 4.1,

$$\lim_{x\to a} x^2 = \lim_{x\to a} x \cdot x = \lim_{x\to a} x \cdot \lim_{x\to a} x = a^2.$$

We can make sure, in a similar way, that

$$\lim_{x\to a} x^n = a^n.$$

Consequently, we shall obtain for the polynomial $b_0x^n + b_1x^{n-1} + \ldots + b_{n-1}x + b_n$ (making use of Theorem 4.1 for a product and a sum)

$$\lim_{x\to a} (b_0x^n + b_1x^{n-1} + \ldots + b_{n-1}x + b_n) = b_0a^n + b_1a^{n-1} + \ldots + b_{n-1}a + b_n.$$

In the case of an irreducible algebraic fraction, when a is not a root of the denominator, we shall obtain (employing Theorem 4.1 for the quotient):

$$\lim_{x\to a} \frac{b_0x^n + b_1x^{n-1} + \ldots + b_{n-1}x + b_n}{c_0x^m + c_1x^{m-1} + \ldots + c_{m-1}x + c_m} = \frac{b_0a^n + b_1a^{n-1} + \ldots + b_{n-1}a + b_n}{c_0a^m + c_1a^{m-1} + \ldots + c_{m-1}a + c_m}.$$

4.2.3. Comparing infinitesimal and infinitely large functions. *A function $y = f(x)$ is said to be infinitesimal at point $x = a$ (for $x \to a$) if* $\lim_{x\to a} f(x) = 0$. It is easy to make sure, for example, that the function $f(x) = (x - a)^m$, where m is a positive integer, is infinitesimal at point $x = a$. Indeed, we have established in the previous subsection that the limit of the polynomial $f(x) = (x - a)^m$ at any point of an infinite number line exists and is equal to a particular value of the polynomial at that point. Therefore, $\lim_{x\to a} (x - a)^m = 0$.

Note that *if the function $y = f(x)$ has a limit equal to b at point a, then the function $\alpha(x) = f(x) - b$ is infinitesimal at point a.* Indeed, the limiting values of each of the functions $f(x)$ and b at point a are equal to b and, therefore, by virtue of Theorem 4.1,

$$\lim_{x\to a} \alpha(x) = \lim_{x\to a} (f(x) - b) = \lim_{x\to a} f(x) - \lim_{x\to a} b = 0.$$

Making use of the result obtained, we get a special representation

* An irreducible algebraic fraction is the quotient of two polynomials which have no common divisors different from a constant.

for a function having a limiting value equal to b at point $x = a$:

$$f(x) = b + \alpha(x), \text{ where } \lim_{x \to a} \alpha(x) = 0. \tag{4.1}$$

Representation (4.1) turns out to be very convenient to prove various propositions and will be repeatedly used in what follows.

Besides the concept of an infinitesimal function, frequent use is made of the concept of an infinitely large function at point a on the right or an infinitely large function at point a on the left. Namely, *the function $f(x)$ is said to be infinitely large at point a on the right (on the left), if for any sequence $x_1, x_2, \ldots, x_n, \ldots$ of the values of the argument x, converging to a, whose terms x_n are greater than a (smaller than a), the corresponding sequence $f(x_1), f(x_2), \ldots, f(x_n), \ldots$ of the values of the function is an infinitely large sequence of a definite sign.*

Infinitely large functions are designated as follows:

$$\lim_{x \to a+0} f(x) = +\infty \quad \text{or} \quad f(a+0) = +\infty,$$

$$\lim_{x \to a-0} f(x) = +\infty \quad \text{or} \quad f(a-0) = +\infty,$$

$$\lim_{x \to a+0} f(x) = -\infty \quad \text{or} \quad f(a+0) = -\infty,$$

$$\lim_{x \to a-0} f(x) = -\infty \quad \text{or} \quad f(a-0) = -\infty.$$

Let us get acquainted with the techniques used to compare infinitesimal functions and the terminology used.

Let $\alpha(x)$ and $\beta(x)$ be two functions defined on the same set and being infinitesimal at a point $x = a$.

(1) The function $\alpha(x)$ is said to be an *infinitesimal function of the order higher than that of* $\beta(x)$ (of a higher order of smallness), if the limiting value of the function $\frac{\alpha(x)}{\beta(x)}$ at point a is equal to zero.

(2) The functions $\alpha(x)$ and $\beta(x)$ are said to be *infinitesimals of the same order* (to have the same order of smallness), if the limiting value of the function $\frac{\alpha(x)}{\beta(x)}$ at point a exists and is different from zero.

(3) The functions $\alpha(x)$ and $\beta(x)$ are said to be *equivalent infinitesimals*, if the limiting value of the function $\frac{\alpha(x)}{\beta(x)}$ at point a is equal to unity.

Infinitesimal functions are often compared to some standard infinitesimal functions. The function $(x - a)^m$ often serves as a standard function. Here m is a positive integer. The following terminology is used in this case: the infinitesimal function $\alpha(x)$ at a point a has the *order of smallness* m if the limiting value of the function $\frac{\alpha(x)}{(x-a)^m}$ at point a is different from zero.

To compare infinitesimal functions, use is often made of the symbol o (lower-case letter), namely, if the function $\alpha = \alpha(x)$ is an infinitesimal function, at point a, of the order higher than that of the infinitesimal function $\beta = \beta(x)$ at that point, it is symbolized as follows:

$$\alpha = o(\beta)$$

(read: a is equal to a lower-case o of β). Thus, the symbol $o(\beta)$ means an *arbitrary* infinitesimal function having, at point a, a higher order of smallness than the infinitesimal function $\beta = \beta(x)$ at that point.

We want to point out the following obvious properties of the symbol o: if $\gamma = o(\beta)$, then $o(\beta) \pm o(\gamma) = o(\beta)$, $o(\beta) \pm o(\beta) = o(\beta)$.

Make a note of the fact that if α and β are infinitesimal functions at point a, then the function $\alpha\beta$ has the order of smallness higher than that of each of the factors, and therefore

$$\alpha\beta = o(\alpha), \quad \alpha\beta = o(\beta).$$

The method of comparison of infinitely large functions on the right (on the left) at point a is quite analogous.

Let $A(x)$ and $B(x)$ be infinitely large, at a point a, functions on the right, and let, for instance, both of them be of the same sign, i.e.

$$\lim_{x \to a+0} A(x) = +\infty \quad \text{and} \quad \lim_{x \to a+0} B(x) = +\infty.$$

We shall say that at point a on the right the function $A(x)$ has a *higher order of growth* than the function $B(x)$, if the function $\frac{A(x)}{B(x)}$ is infinitely large at point a on the right. Now if the right-hand limiting value of the function $\frac{A(x)}{B(x)}$ at point a is finite and nonzero, we shall say that $A(x)$ and $B(x)$ have the *same order of growth* on the right at point a.

Let us consider some examples.

1. The functions $\alpha(x) = 3x^2 + x^3$ and $\beta(x) = 2x^2$ are infinitesimal functions of the same order at point $x = 0$. Indeed, $\frac{\alpha(x)}{\beta(x)} = \frac{3}{2} + \frac{1}{2}x$ when x is different from zero. Since $\lim_{x \to 0} \frac{1}{2}x = 0$, we have, by virtue of Theorem 4.1, $\lim_{x \to 0} \frac{\alpha(x)}{\beta(x)} = \frac{3}{2}$. And this means that $\alpha(x)$ and $\beta(x)$ are infinitesimals of the same order.

2. The functions $\alpha(x) = x^2 - 6x^3$ and $\beta(x) = x^2$ are equivalent infinitesimal functions at point $x = 0$. Indeed, $\frac{\alpha(x)}{\beta(x)} = 1 - 6x$. Since $\lim_{x \to 0} 6x = 0$, we have, by virtue of Theorem 4.1, $\lim_{x \to 0} \frac{\alpha(x)}{\beta(x)} = 1$. And

this precisely means that the infinitesimals $\alpha(x)$ and $\beta(x)$ are equivalent.

3. The functions $A(x)=\frac{1+x}{x}$ and $B(x)=\frac{1}{x}$ have the same order of growth, at point $x=0$, on the right and on the left. This follows from the fact that $\lim\limits_{x\to 0}\frac{A(x)}{B(x)}=\lim\limits_{x\to 0}(1+x)=1$.

4.3. THE CONCEPT OF CONTINUITY OF A FUNCTION

4.3.1. Definition of continuity of a function. Let us assume that point a belongs to the domain of definition of the function $f(x)$ and that any ε-neighbourhood of point a contains points from the domain of definition of that function different from a.

Definition 1. *The function $f(x)$ is said to be* **continuous** *at a point a if the limiting value of that function at point a exists and is equal to the particular value $f(a)$.*

Thus the condition of continuity of the function $f(x)$ at point a can be symbolized as follows:

$$\lim_{x\to a} f(x)=f(a).$$

Since $a=\lim\limits_{x\to a}x$, this equality can be written in the following form:

$$\lim_{x\to a} f(x)=f\left(\lim_{x\to a}x\right).$$

Consequently, for a continuous function the symbol "lim" of a limiting process and the symbol "f" of the characteristic of the function can change places.

Making use of Definition 1 of the limiting value of the function $f(x)$ at point a (see 2.1 of the present chapter), we can reformulate Definition 1 of the continuity of a function at point a as follows.

Definition 1**. *The function $f(x)$ is said to be* **continuous** *at point a if for any, converging to a, sequence $x_1, x_2, \ldots, x_n, \ldots$ of the values of the argument x the corresponding sequence $f(x_1), f(x_2), \ldots, f(x_n), \ldots$ of the values of that function converges to the number $f(a)$.*

Note that as compared to Definition 1 of the limiting value of $f(x)$ at point a presented in 4.2.1, we have omitted in Definition 1 the requirement that all the terms of the sequence $x_1, x_2, \ldots, x_n, \ldots$ should be different from a. This omission is legitimate since the addition of any number of new terms equal to $f(a)$ to those of the sequence $\{f(x_n)\}$ converging to $f(a)$ will not violate the convergence of the resulting sequence to $f(a)$.

Let us postulate that the set $\{x\}$ on which the function $f(x)$ is defined contains a point a and that for any $\varepsilon > 0$ there is at least one member of that set lying on the interval $(a, a + \varepsilon)$ (on the interval $(a - \varepsilon, a)$).

Definition 2. *The function $f(x)$ is called* **continuous on the right (on the left)** *at point a if the right-hand (left-hand) limiting value of that function at point a exists and is equal to the particular value $f(a)$.*

The notation of the continuity on the right (on the left):

$$\lim_{x \to a+0} f(x) = f(a) \quad \text{or} \quad f(a+0) = f(a),$$

$$\left(\lim_{x \to a-0} f(x) = f(a) \quad \text{or} \quad f(a-0) = f(a)\right).$$

Remark. *If the function $f(x)$ is continuous at point a both on the right and on the left, it is continuous at that point.* Indeed, in accordance with the remark made in 2.1 of this chapter, there exists in this case a limiting value of the function at point a equal to a particular value of that function at point a.

Consider some examples.

1°. The power function $f(x) = x^n$ with the integral positive exponent n is continuous at every point of an infinite number line. Indeed, we have proved in 4.2.2 that the limiting value of that function at any point of an infinite number line is equal to the particular value a^n.

2°. Since polynomials and irreducible algebraic fractions have, at every point of the domain of definition, a limiting value equal to a particular value (see 4.2.2), they are continuous functions.

The points at which a function does not possess the property of continuity are called *points of discontinuity** of a function. For instance, the function $f(x) = \operatorname{sgn} x$ has a discontinuity at point $x = 0$ (we have proved in 4.2.1 that the right-hand and left-hand limits of that function at point $x = 0$ exist but are not equal to each other and, therefore, a limiting value of the function does not exist at that point). The Dirichlet function has a discontinuity at each point of an infinite interval since it does not have a limiting value at any point of that interval (see 4.2.1).

We shall say that the function $f(x)$ is *continuous on the set* $\{x\}$ if it is continuous at every point of that set. If a function is continuous at every point of an interval, it is said to be continuous on that interval. If a function is continuous at every interior point of the closed interval $[a, b]$, and, besides, is continuous on the right at point a and on the left at point b, it is said to be continuous on the interval $[a, b]$.

* Points of discontinuity will be classified in 4.8

4.3.2. Arithmetic operations on continuous functions. We shall now make sure that arithmetic operations on continuous functions lead to continuous functions.

Let us prove the following *fundamental* theorem.

Theorem 4.2. *Suppose that the functions $f(x)$ and $g(x)$ defined on the same set are continuous at point a. Then the functions $f(x) + g(x)$, $f(x) - g(x)$, $f(x) \cdot g(x)$ and $\frac{f(x)}{g(x)}$ are continuous at point a (a quotient under the condition $g(a) \neq 0$).*

Proof. Since the functions $f(x)$ and $g(x)$ continuous at point a have the limiting values $f(a)$ and $g(a)$ at that point, in accordance with Theorem 4.1 the limiting values of the functions $f(x) + g(x)$, $f(x) - g(x)$, $f(x) \cdot g(x)$ and $\frac{f(x)}{g(x)}$ exist and are equal, respectively, to $f(a) + g(a)$, $f(a) - g(a)$, $f(a) \cdot g(a)$ and $\frac{f(a)}{g(a)}$. But these values are precisely equal to the particular values of the indicated functions at point a. We have proved the theorem.

4.3.3. Composite function and its continuity. Functions obtained as a result of superposition (i.e. consecutive application) of two or several functions are called composite.

It is sufficient to define a composite function appearing as a result of superposition of two functions.

Suppose the function $x = \varphi(t)$ is defined on a certain set $\{t\}$ and assume that $\{x\}$ is the set of values of that function.

Assume, furthermore, that another function $y = f(x)$ is defined on the indicated set $\{x\}$. Then it is said that defined on the set $\{t\}$ is the composite function

$$y = f(x), \text{ where } x = \varphi(t),$$

or

$$y = f[\varphi(t)] = F(t).$$

The following *fundamental* theorem is valid.

Theorem 4.3. *If the function $x = \varphi(t)$ is continuous at point a and the function $y = f(x)$ is continuous at the corresponding point $b = \varphi(a)$, then the composite function $y = f[\varphi(t)] = F(t)$ is continuous at point a.*

Proof. Assume that $\{t_n\}$ is an arbitrary sequence of the values of the argument of a composite function converging to a. The function $x = \varphi(t)$ being continuous at point a, Definition 1* presented in 4.3.1 allows the corresponding sequence of the values of that function $x_n = \varphi(t_n)$ to converge to the particular value of that function at point a, that is, to the number $b = \varphi(a)$. To proceed further on, since the function $y = f(x)$ is continuous at point $b = \varphi(a)$ and, for this function, the indicated sequence $\{x_n\}$, converging to $b = \varphi(a)$, is the sequence of values of the argument, the corresponding sequence

of values of the function $f(x_n) = f[\varphi(t_n)] = F(t_n)$ converges to the number $f(b) = f[\varphi(a)] = F(a)$ by virtue of the same Definition 1* from 4.3.1.

Thus we see that for any sequence $\{t_n\}$ of values of the argument of a composite function converging to a a corresponding sequence $\{f[\varphi(t_n)]\} \equiv \{F(t_n)\}$ of values of the composite function itself converges to the number $f[\varphi(a)] = F(a)$, which is a particular value of the composite function at point a. By virtue of the same Definition 1* given in 4.3.1, this means that the composite function $f[\varphi(t)] = F(t)$ is continuous at point a. We have proved the theorem.

4.4. SOME PROPERTIES OF MONOTONIC FUNCTIONS

4.4.1. Definition and examples of monotonic functions.

Definition. *The function $y = f(x)$ is called* **nondecreasing (nonincreasing) on the set $\{x\}$,** *if for any x_1 and x_2 of that set satisfying the condition $x_1 < x_2$ there holds an inequality $f(x_1) \leqslant f(x_2)$ ($f(x_1) \geqslant f(x_2)$).*

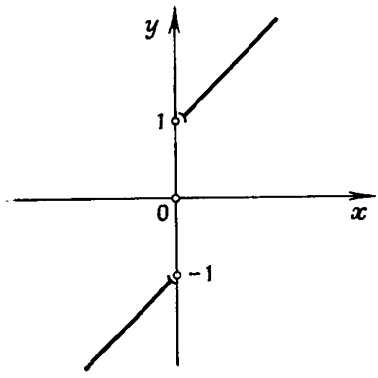


Fig. 4.7

Nondecreasing and nonincreasing functions have a common name *monotonic (monotone)* functions.

If the inequality $f(x_1) < f(x_2)$ ($f(x_1) > f(x_2)$) is valid for any x_1 and x_2 from the set $\{x\}$ satisfying the condition $x_1 < x_2$, then the function $y = f(x)$ is said to be *increasing (decreasing)* on the set $\{x\}$. Increasing and decreasing functions are also called *strictly monotonic.*

Here are some examples of monotonic functions.

1. The function $f(x) = x + \operatorname{sgn} x$ increases on the whole number line (Fig. 4.7).

2. The function $f(x) = \operatorname{sgn} x$ is nondecreasing on the whole number line (see Fig. 4.4).

4.4.2. The concept of an inverse function. A monotonic function having an inverse. Here we shall formulate the concept of an inverse function and establish the conditions of the existence of an inverse function for a monotone function.

Let the function $y = f(x)$ be defined on the closed interval $[a, b]$ and let the closed interval $[\alpha, \beta]$ be the set of values of that function. Now let each value y from the interval $[\alpha, \beta]$ be associated with only one value x of the interval $[a, b]$, for which $f(x) = y$. Then we can define the function $x = f^{-1}(y)$ on the interval $[\alpha, \beta]$ by associating every y from $[\alpha, \beta]$ with the value of x from $[a, b]$ for which $f(x) = y$. The function $x = f^{-1}(y)$ is called the **inverse** *of the function $y = f(x)$.*

In the definition made we could consider the *open intervals* (a, b) and (α, β) *instead of the closed intervals* $[a, b]$ and $[\alpha, \beta]$. We could also assume that one of the intervals (a, b) and (α, β) or both of them turn into an infinite interval or into an open half-line.

It should be noted that if $x = f^{-1}(y)$ is an inverse of the function $y = f(x)$, then, evidently, the function $y = f(x)$ is an inverse of the function $x = f^{-1}(y)$. Therefore, the functions $y = f(x)$ and $x = f^{-1}(y)$ are also called *mutually inverse.*

Mutually inverse functions have the following obvious properties:

$$f(f^{-1}(y)) = y, \quad f^{-1}(f(x)) = x.$$

Let us consider some examples of mutually inverse functions.

1°. Let the function $f(x) = 3x$ be defined on the closed interval $[0, 1]$. The interval $[0, 3]$ is the set of values of that function. The function $f^{-1}(y) = \frac{1}{3}y$ defined on the interval $[0, 3]$ is the inverse of the given function $f(x) = 3x$.

2°. Let us consider, on the closed interval $[0, 1]$, a function defined as follows:

$$y = f(x) = \begin{cases} x & \text{if } x \text{ is a rational number,} \\ 1 - x & \text{if } x \text{ is an irrational number.} \end{cases}$$

The function $x = f^{-1}(y)$ defined on the closed interval $[0, 1]$ and determined by the equalities

$$x = f^{-1}(y) = \begin{cases} y & \text{if } y \text{ is a rational number,} \\ 1 - y & \text{if } y \text{ is an irrational number,} \end{cases}$$

is the inverse of the original function. This fact can be directly verified.

Remark 1. Let the strictly monotone function $y = f(x)$ be defined on the closed interval $[a, b]$ and let the closed interval $[\alpha, \beta]$ be the set of values of that function. Then, by virtue of the strict monotonicity of the function $y = f(x)$, every y belonging to $[\alpha, \beta]$ is associated with *only one* value x belonging to $[a, b]$, for which $f(x) = y$, and therefore there exists a function $x = f^{-1}(y)$ on the interval $[\alpha, \beta]$ which is the inverse of $y = f(x)$. Moreover, if the function $y = f(x)$ is increasing on the interval $[a, b]$, then the function $x = f^{-1}(y)$ is also increasing on the interval $[\alpha, \beta]$; now if the function $y = f(x)$ is decreasing on $[a, b]$, then $x = f^{-1}(y)$ is a decreasing function on the interval $[\beta, \alpha]$. Let us make sure, for instance, that if $y = f(x)$ is an increasing function, then $x = f^{-1}(y)$ is an increasing function as well. Indeed, if $y_1 < y_2$, then $x_1 < x_2$ as well ($x_1 = f^{-1}(y_1)$ and $x_2 = f^{-1}(y_2)$), since the inequality $x_1 \geqslant x_2$ and the increase of the function $y = f(x)$ would yield $y_1 \geqslant y_2$ and this would contradict the inequality $y_1 < y_2$.

Lemma 1. *For the function $y = f(x)$ strictly monotonic on the closed interval $[a, b]$ to be continuous on that interval it is necessary and sufficient that any number γ contained between the numbers $\alpha = f(a)$ and $\beta = f(b)$ be a value of that function.* In other words, for the strictly monotonic function $y = f(x)$ to be continuous on the closed interval $[a, b]$, it is necessary and sufficient that the interval $[\alpha, \beta]$ (or $[\beta, \alpha]$ with $\beta < \alpha$) be the set of values of that function. Here $\alpha = f(a)$ and $\beta = f(b)$.

Proof. (1) **Necessity.** For definiteness, let us consider an increasing function $y = f(x)$ continuous on the closed interval $[a, b]$ (for a decreasing function the proof is quite analogous). We shall show that if $\alpha < \gamma < \beta$ then there exists an interior point c of the interval $[a, b]$ at which $f(c) = \gamma$ (by virtue of the fact that the function $f(x)$ is increasing on the interval $[a, b]$ such a point c is unique). Let us denote by $\{x\}$ the set of points belonging to the interval $[a, b]$ for which $f(x) \leqslant \gamma$ (this set contains point a, for instance, since $f(a) = \alpha < \gamma$). The set $\{x\}$ is bounded above and consequently it has the least upper bound c. We must prove that $f(c) = \gamma$. Note that any number, belonging to the interval $[a, b]$, which is smaller than c, is contained in the set $\{x\}$* while any number exceeding c does not belong to that set**. We shall show that c is an interior point of the closed interval $[a, b]$. Indeed, let c be equal to b, for instance. We shall consider an increasing sequence $\{x_n\}$ of values of the argument of the function $y = f(x)$ converging to b. Since the function $f(x)$ is continuous on the left at point b, we have $\lim_{n\to\infty} f(x_n) = \beta$. On the other hand, $f(x_n) \leqslant \gamma$***, and, therefore, in accordance with Theorem 3.13, $\lim_{n\to\infty} f(x_n) \leqslant \gamma$. Thus, $\beta \leqslant \gamma$, but that contradicts the hypothesis $\gamma < \beta$. The contradiction obtained proves that $c < b$. It can be proved by analogy that $a < c$. Since c is an interior point of the interval $[a, b]$, there can be found $\{x'_n\}$ and $\{x''_n\}$, an increasing and a decreasing sequence of values of the argument x converging to c. Since the function $f(x)$ is continuous at point c, we have $\lim_{n\to\infty} f(x'_n) = \lim_{n\to\infty} f(x''_n) = f(c)$. But $f(x'_n) \leqslant \gamma$ and $f(x''_n) > \gamma$****. Therefore, $\lim_{n\to\infty} f(x'_n) \leqslant \gamma$, $\lim_{n\to\infty} f(x''_n) \geqslant \gamma$, whence $f(c) = \gamma$.

(2) **Sufficiency.** We shall now present the proof for the function $y = f(x)$ increasing on the closed interval $[a, b]$ (the argument for

* For, according to the definition of the least upper bound, there can be found, for any $x < c$, an x' such that $x < x'$ and $f(x') \leqslant \gamma$. But then it follows from an increase of the function $f(x)$ that $f(x) \leqslant \gamma$ as well, i.e. that x belongs to $\{x\}$.

** By virtue of the definition of the least upper bound.

*** Since all x_n are smaller than c and therefore belong to $\{x\}$.

**** By virtue of the fact that $x'_n < c < x''_n$ for any number n.

a decreasing function is analogous). Assume that c is an arbitrary point belonging to the interval $[a, b]$ and $\gamma = f(c)$ is the value of the function $y = f(x)$ at that point. Let us make sure that the number γ is the right-hand and the left-hand limit of the function $f(x)$ at point c (if c is the boundary point of the interval $[a, b]$, then γ is the corresponding one-sided limiting value at that boundary point). Let $a < c \leqslant b$; we shall prove that γ is the left-hand limiting value of the function at point c. Suppose that ε is a small integer such that $\alpha < \gamma - \varepsilon$ (Fig. 4.8). Since by the condition of the lemma the number $\gamma - \varepsilon$ is a value of the function $f(x)$, there can be indicated, on the interval $[a, b]$, a point d such that $f(d) = \gamma - \varepsilon$. Inasmuch as the function $f(x)$ increases, we have $d < c$. Let us now consider an arbitrary sequence $\{x_n\}$, converging to c, of the values of the argument x, whose terms are smaller than c. Beginning with some number N, all the terms x_n of that sequence satisfy the inequalities $d < x_n < c$ (one of such terms is shown in Fig. 4.8), so that because of the increase of the function $f(x)$ for $n \geqslant N$ the inequalities $f(d) < f(x_n) < f(c)$ are valid. Since $f(d) = \gamma - \varepsilon$ and $f(c) = \gamma$, it follows from the last inequalities that the inequalities $0 < \gamma - f(x_n) < \varepsilon$ are valid when $n \geqslant N$. In other words, the sequence $\{f(x_n)\}$ converges to γ, and since $\{x_n\}$ is an arbitrary sequence of values of the argument converging to c on the left, this is the proof that the left-hand limiting value at point c exists and is equal to $\gamma = f(c)$*. If $a \leqslant c < b$, then an analogous argument enables us to prove that $\gamma = f(c)$ is the right-hand limiting value of the function at point c. We have proved that the right-hand and left-hand limiting values of the function $y = f(x)$ at any interior point c are equal to its particular value $f(c)$, and this, in accordance with the remark in 4.2.1, implies the continuity of $f(x)$ at interior points of the interval. The continuity of that function at the boundary points of the interval follows from the fact that the corresponding one-sided limiting values of $f(x)$ at the boundary points of the interval are equal to the particular values of the function. Thus we have completed the proof of the lemma.

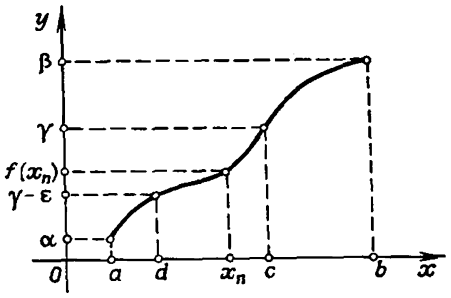


Fig. 4.8

Corollary. *Let a strictly monotone continuous function $y = f(x)$ be defined on the closed interval $[a, b]$, and let $\alpha = f(a)$, $\beta = f(b)$. Then*

* We have considered the case of a positive ε so small that $\alpha < \gamma - \varepsilon$. Now if $\alpha \geqslant \gamma - \varepsilon$, it is sufficient to put $d = a$ and repeat the argument, making use of the obvious inequality $\gamma - \varepsilon \leqslant f(d)$.

this function has an inverse strictly monotone and continuous function $x = f^{-1}(y)$ *on the closed interval* $[\alpha, \beta]$ (or $[\beta, \alpha]$ if $\beta < \alpha$).

Proof. By virtue of the lemma just proved, the set of values of the function $y = f(x)$ is the closed interval $[\alpha, \beta]$. Then, in accordance with Remark 1 made in this subsection, there exists an inverse strictly monotone function $x = f^{-1}(y)$ on the interval $[\alpha, \beta]$ whose set of values is the interval $[a, b]$ and which, therefore, is continuous on the interval $[\alpha, \beta]$ by virtue of the same lemma.

Remark 2. It should be noted that monotonic functions have both the right-hand and the left-hand limiting values at each interior point of the domain of definition. The proof of this proposition is left to the reader.

4.5. BASIC ELEMENTARY FUNCTIONS

The class of basic elementary functions includes the following functions: $y = x^{\alpha}$, $y = a^{x}$, $y = \log_a x$, $y = \sin x$, $y = \cos x$, $y = \tan x$, $y = \cot x$, $y = \arcsin x$, $y = \arccos x$, $y = \arctan x$, $y = \operatorname{arccot} x$.

These functions and their graphs are familiar from the course of elementary mathematics studied at school. Some of them, for instance $y = a^{x}$, can be easily determined for rational values of the argument x. We shall elucidate the question of determining the basic elementary functions for various real values of their arguments. This is not an easy question: it is not clear, for instance, how to raise an arbitrary real number x to an arbitrary real power α.

We shall also discuss the continuity of the basic elementary functions at all points of their domain of definition. We shall substantiate the behaviour of the basic elementary functions which is evident from their graphs.

The supplement to Chapter 8 includes the algorithms for computing the values of basic elementary functions.

4.5.1. Rational powers of positive numbers. The raising of any real number x to the *positive integral power* n is defined as an n-fold multiplication of the number x by itself. Consequently, for an integral n, we shall assume the power function $y = x^{n}$ to be defined for all the real values of x. Some properties of this function will be employed in the process of defining rational powers of positive numbers.

Let us prove the following lemma.

Lemma 2. *The power function* $y = x^{n}$ *for* $x \geqslant 0$ *and a positive integral* n *increases and is continuous.*

Proof. Let us show that the function is increasing. Assume that $0 \leqslant x_1 < x_2$. Since $x_2^{n} - x_1^{n} = (x_2 - x_1)(x_2^{n-1} + x_2^{n-2}x_1 + \ldots + x_1^{n-1})$ and $x_2^{n-1} + x_2^{n-2}x_1 + \ldots + x_1^{n-1} > 0$, it follows that

$x_2^n > x_1^n$. The continuity of this function was established earlier (see Example 1 in 4.3.1).

Corollary. Let us consider the power function $y = x^n$ on the interval $[0, N]$, where N is an arbitrary positive number. Since this function is continuous and increases on the indicated interval, it has, in accordance with the corollary of Lemma 1 of this chapter, an increasing and continuous inverse function on the closed interval $[0, N^n]$ which we shall denote by $y^{1/n}$.

Since N can be taken arbitrarily large, N^n will also be arbitrarily large. Consequently, the function $x = y^{1/n}$ is defined for all the nonnegative values of y. Replacing the designation y of the argument of this function by x and the designation of x of the function itself by y, we shall obtain a power function $y = x^{1/n}$ defined for all the nonnegative values of x.

Let us define $a^{1/n}$ as the number b equal to the value of the function $y = x^{1/n}$ at point a. We are now able to determine any rational power r of the positive number a. Namely, if $r = m/n$, where m and n are positive integers, we shall put

$$a^r = a^{m/n} = (a^{1/n})^m.$$

We shall assume in addition that

$$a^0 = 1, \quad a^{-r} = \left(\frac{1}{a}\right)^r.$$

It is easy to verify the validity of the following properties of the rational power of positive numbers:

$$(a^r)^s = a^{r \cdot s}, \quad a^r \cdot b^r = (a \cdot b)^r, \quad a^r \cdot a^s = a^{r+s}. \qquad (*)$$

Let us first prove the validity of the first property (*). Note that for an integral positive p the equality $(a^{m/n})^p = a^{\frac{m \cdot p}{n}}$, in which m and n are understood as any integral positive numbers, is obviously valid since both the left-hand side and the right-hand side of this equality are equal to the product of the number $a^{1/n}$ multiplied by itself $m \cdot p$ times.

Putting $r = \frac{m_1}{n_1}$, $s = \frac{m_2}{n_2}$, we shall prove the equality $(a^r)^s = a^{r \cdot s}$, r and s being any positive rational numbers. Let us put $c_1 = (a^{m_1/n_1})^{m_2/n_2}$, $c_2 = a^{\frac{m_1 \cdot m_2}{n_1 \cdot n_2}}$. Were c_1 different from c_2, it would follow from an increase of the power function $y = x^{n_2}$ that $c_1^{n_2} \neq c_2^{n_2}$ as well, and the last relation would mean, by virtue of the validity of the equality $(a^{m/n})^p = a^{\frac{m \cdot p}{n}}$ for an integral p we have proved, that $(a^{m_1/n_1})^{m_2} \neq a^{\frac{m_1 \cdot m_2}{n_1}}$. The resulting equation contradicts the equality $(a^{m_1/n_1})^{m_2} = a^{\frac{m_1 \cdot m_2}{n_1}}$ we have proved for integral positive m_1, n_1 and m_2. Thus we infer that $c_1 = c_2$ and consider the first equality (*) to be proved for any positive rational r and s.

It is not difficult to extend this equality to nonpositive r and s since we have agreed that

$$a^0 = 1, \quad a^{-r} = \left(\frac{1}{a}\right)^r \text{ for } r > 0.$$

As to the second equality (*), it is also sufficient to prove it for *positive* rational r. Assuming this r to be equal to $\frac{m}{n}$, where m and n are positive integers, we note that it suffices to prove the equality $a^{1/n} \cdot b^{1/n} = (a \cdot b)^{1/n}$, since by multiplying m equalities of this kind, we can prove the general relation $a^r \cdot b^r = (a \cdot b)^r$.

To prove the equality $a^{1/n} \cdot b^{1/n} = (a \cdot b)^{1/n}$, pay attention to the fact that by virtue of the properties of mutually inverse functions $y = x^{1/n}$ and $x = y^n$, we can assert that $(b^{1/n})^n = b$, $(a^{1/n})^n = a$, $((a \cdot b)^{1/n})^n = a \cdot b$. Therefore, putting $c_1 = a^{1/n} \cdot b^{1/n}$, $c_2 = (a \cdot b)^{1/n}$ and assuming that $c_1 \neq c_2$, we should have received $c_1^n \neq c_2^n$ which contradicts the equality $a \cdot b = ab$.

Let us now prove the last property (*), taking into account that the first two properties have been proved. Assume $r = \frac{m_1}{n_1}$, $s = \frac{m_2}{n_2}$, then we have $r = \frac{m_1 \cdot n_2}{n_1 n_2}$, $s = \frac{m_2 \cdot n_1}{n_2 \cdot n_1}$ and we arrive at the following equality:

$$a^r \cdot a^s = \left(a^{\frac{1}{n_1 n_2}}\right)^{m_1 n_2} \cdot \left(a^{\frac{1}{n_1 n_2}}\right)^{m_2 n_1} = \left(a^{\frac{1}{n_1 n_2}}\right)^{m_1 n_2 + m_2 n_1}.$$

The last equality holds true since $m_1 n_2$ and $m_2 n_1$ are integers.

Thus we have

$$a^r \cdot a^s = a^{\frac{m_1 n_2 + m_2 n_1}{n_1 n_2}} = a^{\frac{m_1}{n_1} + \frac{m_2}{n_2}} = a^{r+s},$$

and this is what we had to prove.

Let us prove that for $a > 1$ and a rational $r > 0$ there holds an inequality $a^r > 1$. Indeed, suppose that $r = \frac{m}{n}$ and $a^r = a^{m/n} \leqslant 1$. Multiplying term-by-term the n indicated inequalities, we obtain $a^m \leqslant 1$. The last inequality contradicts the inequality $a^m > 1$ obtained by the termwise multiplication of the m inequalities of the type $a > 1$. We shall also make note of the fact that if the rational fraction $r = \frac{m}{n}$ has an odd denominator n, the definition of the rational power can be extended to cover the negative numbers as well, assuming that

$(-a)^r = a^r$ if m is even,

and

$(-a)^r = -a^r$ if m is odd.

4.5.2. Exponential function. From the argument of the preceding subsection it follows that if a is a positive number, then the function $y = a^x$ is defined for all rational x.

It is easy to verify that the function $y = a^x$, $a > 1$, defined on the set $\{x\}$ of all rational numbers, increases monotonically on that set.

Indeed, let x_1 and x_2 be any two rational numbers satisfying the condition $x_2 > x_1$. Then

$$a^{x_2} - a^{x_1} = a^{x_1}(a^{x_2-x_1} - 1).$$

Since $x_2 - x_1 > 0$ and $a > 1$, it follows that $a^{x_2-x_1} > 1$, that is the right-hand side of the last equality is positive, and, therefore, $a^{x_2} > a^{x_1}$. We have proved the increase of the function a^x on the set of rational numbers.

Let us now define the function a^x *on the set of all real numbers.*

We shall fix an *arbitrary real number* x and consider all possible rational numbers α and β satisfying the inequalities

$$\alpha < x < \beta. \tag{4.2}$$

The function a^x will be defined, for $a > 1$, as a real number y complying with the inequalities

$$a^\alpha \leqslant y \leqslant a^\beta. \tag{4.3}$$

Below we shall prove that such a number y *exists and that it is unique.* We shall also prove that the function $y = a^x$ defined possesses the following important properties: (1) *it increases on the whole infinite number line*, (2) *it is continuous at any point x of that line.*

1°. We shall prove, first of all, that for any fixed x and any rational numbers α and β satisfying inequalities (4.2), *there exists a real number y complying with inequalities* (4.3). We shall begin with fixing a rational number β satisfying the right-hand inequalities (4.2) and then consider all possible rational numbers α satisfying the left-hand inequalities (4.2).

Inasmuch as $\alpha < \beta$ and the exponential function defined on the set of rational numbers increases, we have $a^\alpha < a^\beta$. Thus the set $\{a^\alpha\}$ is bounded above and the number a^β is one of the upper bounds of that set. If follows that the set has the least upper bound which will be denoted by y. It remains to prove that y satisfies inequalities (4.3). It follows from the definition of the least upper bound that the left-hand inequality (4.3) is valid, while the validity of the right-hand inequality (4.3) follows from the fact that a^β is one of the upper bounds and y is the least upper bound of the set $\{a^\alpha\}$.

2°. Let us now establish the fact that *there is only one real number y satisfying inequalities* (4.3).

It is sufficient to prove that for any positive ε there are rational numbers α and β, satisfying inequalities (4.2), for which $a^\beta - a^\alpha < \varepsilon$. In fact, in that case any two numbers y_1 and y_2 satisfying inequalities (4.3) must coincide, for the difference between them, taken absolutely, is smaller than any preassigned positive number ε.

Having fixed an arbitrary positive number ε, and a certain rational number β_0 satisfying the right-hand inequality (4.2), we obtain, since $a^\alpha < a^{\beta_0}$,

$$a^\beta - a^\alpha = a^\alpha(a^{\beta-\alpha} - 1) < a^{\beta_0}(a^{\beta-\alpha} - 1).$$

The inequality $a^\beta - a^\alpha < \varepsilon$ will be proved if we establish the possibility of choosing such α and β that $a^{\beta-\alpha} - 1 < \varepsilon/a^{\beta_0}$.

It follows from the argument carried out in Chapter 2 that for any natural number n there can be chosen rational numbers α and β satisfying inequalities

(4.2) so that the difference $\beta - \alpha$ will be smaller than $\frac{1}{n}$. Thus we see that it is sufficient to prove the existence of a natural n for which

$$a^{1/n} - 1 < \frac{\varepsilon}{a^{\beta_0}}. \tag{4.4}$$

Let us verify the possibility of the choice of such a natural n. Let

$$a^{1/n} = 1 + \delta_n.$$

Since $a^{1/n} > 1$, δ_n is positive. Making use of Newton's binomial formula, we shall have $a = (a^{1/n})^n = (1 + \delta_n)^n = 1 + n\delta_n +$ (positive members) $> 1 + n\delta_n$. Hence, $a - 1 > n\delta_n$ and $0 < \delta_n < \frac{a-1}{n}$. Thus it follows that $a^{1/n} - 1 = \delta_n < \frac{a-1}{n}$. Inequality (4.4) will be valid if we take n satisfying the requirement $\frac{a-1}{n} < \frac{\varepsilon}{a^{\beta n}}$ or $n > \frac{(a-1)a^{\beta_0}}{\varepsilon}$. Thus we have proved the uniqueness of the definition of the number y satisfying inequalities (4.3).

Make note of the fact that if x is a rational number and a^x is the value, at point x, of the exponential function previously defined only on the set of rational numbers, then a^x is precisely that unique real number y which satisfies inequalities (4.3).

3°. Let us prove now that the function a^x we have constructed for $a > 1$ *increases on the whole infinite number line.*

Assume x_1 and x_2 to be any real numbers complying with the inequality $x_1 < x_2$. There are, evidently, rational numbers α and β satisfying the inequalities $x_1 < \alpha < \beta < x_2$ (see the assertion proved at the end of 2.2.1). The definition of an exponential function and its increase on the set of rational numbers yield the inequalities $a^{x_1} \leqslant a^{\alpha} < a^{\beta} \leqslant a^{x_2}$, i.e. $a^{x_1} < a^{x_2}$. We have proved that the function a^x increases.

4°. It remains to prove the *continuity* of the function a^x, we have constructed, at *any point* x of an infinite number line.

Let $\{x_n\}$ be any sequence of real numbers converging to x. It is sufficient to show that for any positive ε there is a number N such that for $n \geqslant N$ the inequality $|a^{x_n} - a^x| < \varepsilon$ holds true.

We shall fix an arbitrary $\varepsilon > 0$ and choose rational numbers α and β satisfying inequalities (4.2), so that there holds the inequality $a^{\beta} - a^{\alpha} < \varepsilon$ (the possibility of such a choice of α and β was proved in 2°). Since the sequence $\{x_n\}$ converges to x and $\alpha < x < \beta$, there is a number N such that for $n \geqslant N$ the inequalities $\alpha < x_n < \beta$ hold true. It follows from the inequalities $\alpha < x < \beta$ and $\alpha < x_n < \beta$ and from the property of monotonicity of an exponential function that $a^{\alpha} < a^{x_n} < a^{\beta}$ and $a^{\alpha} < a^x < a^{\beta}$ (for $n \geqslant N$). Since the difference between the numbers a^{β} and a^{α} is smaller than ε and both numbers a^x and a^{x_n} are contained between a^{α} and a^{β}, it follows that $|a^{x_n} - a^x| < \varepsilon$ (for $n \geqslant N$). The proof of the continuity property is completed.

Remark 1. If $0 < a < 1$, then $a = \frac{1}{b}$, where $b > 1$. Then, for $0 < a < 1$, the function $y = a^x$ can be specified as a function $y = b^{-x}$, $b > 1$.

Let us establish some properties of the exponential function $y = a^x$, $a > 1$.

(1) All the values of an exponential function are positive. Indeed, let x be an arbitrary point of the number scale and x', a rational point such that $x' < x$. Since by definition $a^{x'} > 0$ and $a^{x'} < a^x$, it follows that $a^x > 0$.

(2) $\lim\limits_{x \to -\infty} a^x = 0$, $\lim\limits_{x \to +\infty} a^x = +\infty$.

In fact, since $a > 1$, it follows that $a = 1 + \alpha$, where $\alpha > 0$ and $a^n = (1 + \alpha)^n > 1 + n\alpha$. Consequently, $\lim\limits_{n \to +\infty} a^n = +\infty$. By virtue of monotonicity of the function $\lim\limits_{x \to +\infty} a^x = +\infty$. Since $a^{-n} = 1/a^n$, we have $\lim\limits_{n \to \infty} a^{-n} = 0$ and therefore $\lim\limits_{x \to -\infty} a^x = 0$.

(3) From properties (1) and (2) as well as from the monotonicity and continuity of the function $y = a^x$ it follows, in accordance with Lemma 1, that *the values of that function fill in the entire positive half-line* $y > 0$.

(4) For any real numbers x_1 and x_2 there hold relations

$$(a^{x_1})^{x_2} = a^{x_1 x_2}, \quad a^{x_1} b^{x_1} = (a \cdot b)^{x_1}, \quad a^{x_1} a^{x_2} = a^{x_1 + x_2}.$$

We have earlier indicated the validity of these relations for rational exponents. To make sure that the relations hold for any exponents, it is sufficient to consider the sequences $\{x_n'\}$ and $\{x_n''\}$ of the rational numbers converging to x_1 and x_2 respectively. Then, for instance, $a^{x_n'} a^{x_n''} = a^{x_n' + x_n''}$. Passing to the limit for $n \to \infty$ and making use of the continuity property of an exponential function, we shall have $a^{x_1} a^{x_2} = a^{x_1 + x_2}$. The validity of the other relations given above can be proved in a similar way.

Remark 2. We have established properties (1)-(4) of the exponential function $y = a^x$ as well as the continuity and the monotone increase of that function on an infinite number line for the case $a > 1$. Make note of the fact that by virtue of Remark 1 the function $y = a^x$ is continuous and monotonically decreases on an infinite number line when $0 < a < 1$. Besides, properties (1), (3) and (4) are retained for that function, while property (2) is modified as follows:

$$\lim_{x \to -\infty} a^x = +\infty, \quad \lim_{x \to +\infty} a^x = 0.$$

Figures 4.9 and 4.10 show the graphs of the exponential function $y = a^x$ for the cases $a > 1$ and $0 < a < 1$.

Remark 3. The property $a^{x_1 + x_2} = a^{x_1} a^{x_2}$ can serve as the basis for the functional definition of the exponential function $y = a^x$. It can be proved that there exists a unique function $f(x)$ defined on the whole infinite interval and satisfying the following requirements:

(1) for any real x_1 and x_2 it satisfies the relation $f(x_1 + x_2) = f(x_1) f(x_2)$;

(2) it complies with the relations $f(0) = 1$, $f(1) = a$, where $a > 0$;

(3) it is continuous for $x = 0$.

The function a^x constructed above is precisely that kind of a function.

4.5.3. Logarithmic function. Let us consider the *arbitrary closed interval* $[c, d]$ of an infinite number line. On that interval the function $y = a^x$ is strictly monotonic and continuous. Therefore,

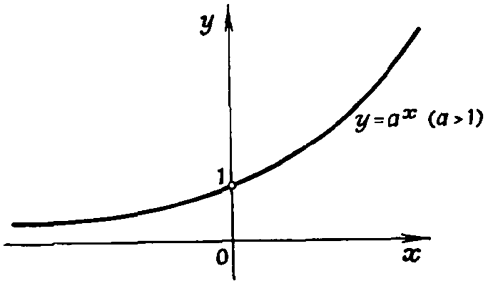


Fig. 4.9

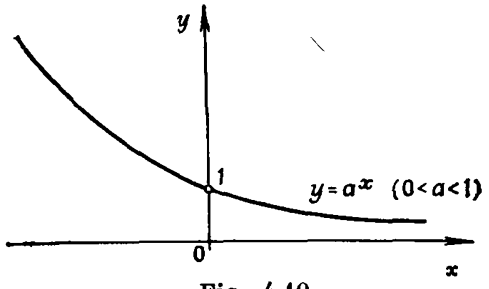


Fig. 4.10

by virtue of the corollary of Lemma 1, the function $y = f(x) = a^x$ possesses on the interval $[\alpha, \beta]$, where $\alpha = a^c$, $\beta = a^d$, an inverse function $x = f^{-1}(y)$ which is called *logarithmic*. A logarithmic function is designated as follows:

$$x = \log_a y.$$

Replacing the designation of the argument y of this function by x and the designation of the function x by y, we shall obtain the function

$$y = \log_a x.$$

Let us note the following properties of a logarithmic function following directly from its definition.

1°. A logarithmic function is defined for all the positive values of x, because its argument constitutes the values of an exponential function which, by virtue of properties (1) and (3) of that function (see previous subsection), can be only positive and fill up the entire positive half-line x.

2°. A logarithmic function is continuous and increases over the entire open half-line $x > 0$ for $a > 1$ (decreases for $a < 1$); in that case, for $a > 1$, we have

$$\lim_{x \to 0+0} \log_a x = -\infty, \quad \lim_{x \to +\infty} \log_a x = +\infty.$$

The validity of this property follows from the properties of an exponential function and from Remark 1 in 4.4.2.

3°. For any positive x_1 and x_2

$$\log_a (x_1 \cdot x_2) = \log_a x_1 + \log_a x_2.$$

This is also a consequence of the properties of an exponential function.

Remark. Special attention should be paid to the logarithmic function $y = \log_e x$, where $e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n$. For this function

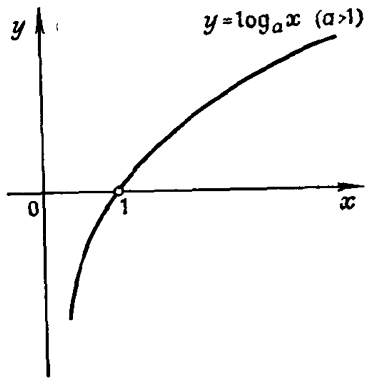


Fig. 4.11

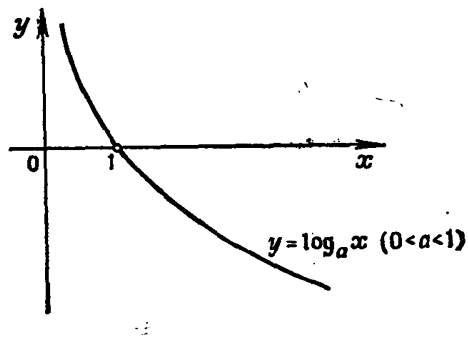


Fig. 4.12

we shall use the notation $y = \ln x$. We wish to emphasize that the logarithmic function $y = \ln x$ plays a very significant role in mathematics and its applications. Logarithms which use the base e are called *natural* logarithms.

Figures 4.11 and 4.12 present the graphs of the logarithmic function $y = \log_a x$ for the cases $a > 1$ and $0 < a < 1$.

4.5.4. Hyperbolic functions. The following functions are called *hyperbolic**:

* The term "hyperbolic functions" can be explained by the fact that the geometric interpretation of the functions $y = \sinh x$ and $y = \cosh x$ follows from the consideration of a rectangular hyperbola in accordance with the same rules which govern the relation of the functions $y = \sin x$ and $y = \cos x$ to a unit circle.

1°. Hyperbolic sine

$$\sinh x = \frac{e^x - e^{-x}}{2}.$$

2°. Hyperbolic cosine

$$\cosh x = \frac{e^x + e^{-x}}{2}.$$

3°. Hyperbolic tangent

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

4°. Hyperbolic cotangent

$$\coth x = \frac{\cosh x}{\sinh x} = \frac{e^x + e^{-x}}{e^x - e^{-x}}.$$

It follows from the definition of hyperbolic functions that a hyperbolic sine, hyperbolic cosine and hyperbolic tangent are defined throughout the number line. A hyperbolic cotangent is defined over the entire number line except for the point $x = 0$.

Hyperbolic functions are continuous at every point of the domain of definition (this follows from the continuity of an exponential function and Theorem 4.2).

Hyperbolic functions possess a number of properties similar to those of trigonometric functions. For instance, the theorems of addition formulated for trigonometric functions hold true for hyperbolic functions as well, namely,

$$\sinh (x + y) = \sinh x \cosh y + \cosh x \sinh y,$$

$$\cosh (x + y) = \cosh x \cosh y + \sinh x \sinh y.$$

Figures 4.13-4.16 illustrate the graphs of hyperbolic functions.

4.5.5. Power function with any real exponent α. Suppose α is an arbitrary real number. Let us define the general *power function* $y = x^\alpha$, $x > 0$ in the following way:

$$y = x^\alpha = (a^{\log_a x})^\alpha = a^{\alpha \log_a x} \quad (a > 1).$$

It follows from the definition of a power function that for $\alpha > 0$ it increases and for $\alpha < 0$ it decreases.

Let us consider a limiting value of the power function for $x \to 0 + 0$. We shall prove that

$$\lim_{x \to 0+0} x^\alpha = \begin{cases} 0 & \text{for } \alpha > 0, \\ +\infty & \text{for } \alpha < 0. \end{cases}$$

Indeed, assume that $\{x_n\}$ is any sequence of values of the argument x converging to zero on the right. Since $\lim_{n \to \infty} \log_a x_n = -\infty$, it

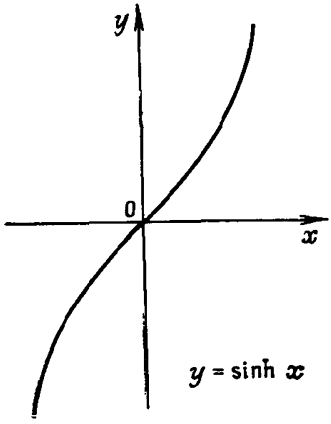


Fig. 4.13

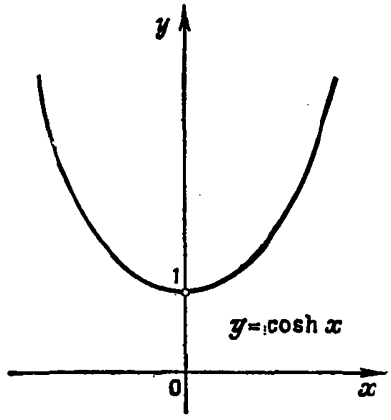


Fig. 4.14

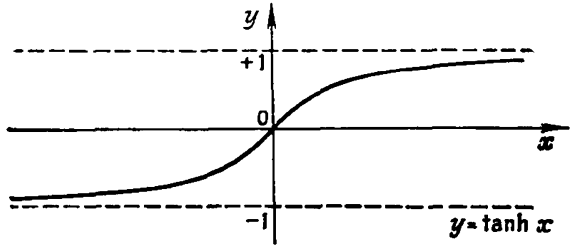


Fig. 4.15

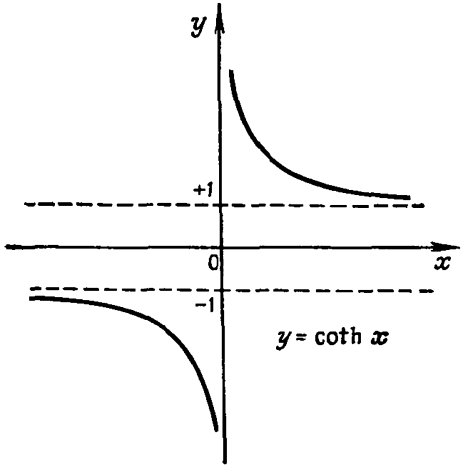


Fig. 4.16

follows from the properties of an exponential function that $\lim\limits_{n\to\infty} a^{\alpha \log_a x_n} = 0$ for $\alpha > 0$ and $\lim\limits_{n\to\infty} a^{\alpha \log_a x_n} = +\infty$ for $\alpha < 0$. It is natural now to put $0^\alpha = 0$ for $\alpha > 0$ and assume that this expression is not defined for $\alpha \leqslant 0$.

We shall now prove the *continuity* of a power function at any point x of a positive infinite half-line ($x > 0$). To do that, it is sufficient to prove that the function is continuous at every point x of the indicated half-line on the left and on the right (see the remark in 4.3.1). We shall establish, for instance, the continuity of that

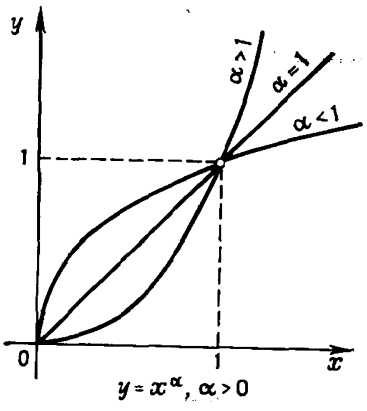


Fig. 4.17

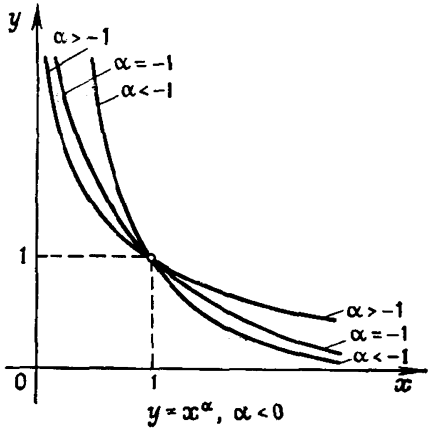


Fig. 4.18

function at point x on the left (continuity on the right can be established analogously). For definiteness, we shall assume that $\alpha > 0$. We shall consider the formula $y = x^\alpha = a^{\alpha \log_a x}$, $a > 1$. Let $\{x_n\}$ be any sequence of values of the argument of a power function converging to x on the left, so that $x_n < x$. Since a logarithmic function is continuous, the sequence $\{u_n\}$, where $u_n = \alpha \log_a x_n$, converges to $u = \alpha \log_a x$, with all the terms u_n being different from u (indeed, since for $a > 1$ the logarithmic function increases, the inequality $u_n < u$ holds true). By virtue of the continuity of an exponential function $\{a^{u_n}\}$ converges to a^u. In other words, the sequence $\{a^{\alpha \log_a x_n}\}$ which is the sequence $\{x_n^\alpha\}$ of the values of the power function corresponding to the sequence $\{x_n\}$, converges to $a^{\alpha \log_a x}$, i.e. to x^α. The continuity of the power function at point $x > 0$ from the left has been proved. The continuity of this function at point $x > 0$ from the right can be proved in the same way. But the continuity of the function at point x on the left and on the right means that the function is continuous at that point. Note that if $\alpha > 0$, the power function $y = x^\alpha$ is continuous at the point $x = 0$ as well.

Remark. We want to point out that if the exponent α of a power function is the rational number $\frac{m}{n}$, where n is an odd integer, we

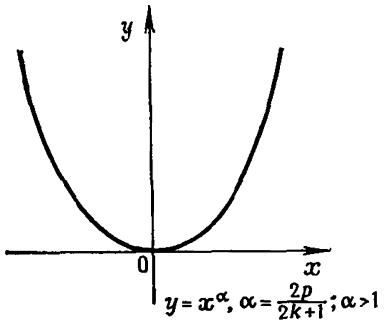


Fig. 4.19

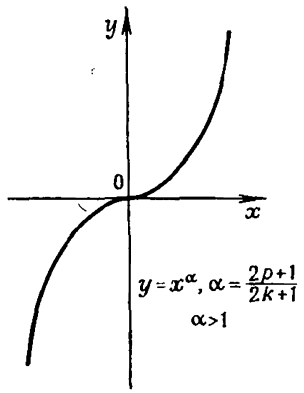


Fig. 4.20

can define the power function $y = x^{\alpha}$ throughout the number line, putting for $x < 0$

$y = |x|^{\alpha}$ if $\alpha = m/n$ and m is even,

$y = -|x|^{\alpha}$ if $\alpha = m/n$ and m is odd.

Figures 4.17-4.20 present the graphs of the power function $y = x^{\alpha}$ for various values of α.

4.5.6. Trigonometric functions. In courses of elementary mathematics we come across the trigonometric functions $y = \sin x$ and $y = \cos x$ considered from the viewpoint of geometry.*

Given below are some properties of trigonometric functions which will turn out to be of importance later on.

1°. For any real x', x'' and x the following relations hold true:

$$\left.\begin{aligned} \sin(x'+x'') &= \sin x' \cos x'' + \cos x' \sin x'', \\ \cos(x'+x'') &= \cos x' \cos x'' - \sin x' \sin x'', \\ \sin^2 x + \cos^2 x &= 1. \end{aligned}\right\} \tag{4.5}$$

* The remaining trigonometric functions such as $y = \tan x$, $y = \cot x$, $y = \sec x$ and $y = \csc x$ can be derived from the functions indicated above:

$$\tan x = \frac{\sin x}{\cos x}, \quad \cot x = \frac{\cos x}{\sin x}, \quad \sec x = \frac{1}{\cos x}, \quad \csc x = \frac{1}{\sin x}.$$

We should like to emphasize that the determination of the functions $\sin x$ and $\cos x$ with the aid of visual geometric means is not logically faultless for in that case the possibility of determining these functions for all the real values of the argument x reduces to the possibility of establishing a one-to-one correspondence between all the points of a unit circle and all the real numbers from the closed interval $[0, 2\pi]$.

2°.

$$\left.\begin{aligned} &\sin 0 = 0;\ \cos 0 = 1, \\ &\sin \frac{\pi}{2} = 1;\ \cos \frac{\pi}{2} = 0. \end{aligned}\right\} \tag{4.6}$$

3°. If $0 < x < \frac{\pi}{2}$, then

$$0 < \sin x < x. \tag{4.7}$$

The indicated properties can be established by means of geometrical reasoning. We shall not give here geometrical derivations of Properties 1° and 2° known from the course of elementary mathematics. We shall only *derive* inequality (4.7) by *geometrical methods*. In addition to (4.7) we shall establish the inequality $x < \tan x$ $\left(\text{for } 0 < x < \frac{\pi}{2}\right)$.

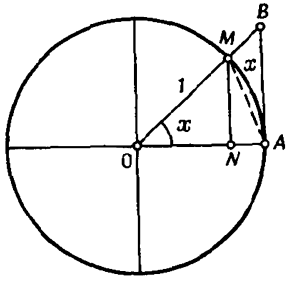


Fig. 4.21

Let us discuss a circle of radius 1 with centre at point O and point A on that circle (Fig. 4.21). We shall reckon the arcs of the circle from point A counterclockwise. Assume that M is a point of the circle located in the first quadrant and x is the length of the arc $\overset{\frown}{AM}$, $0 < x < \frac{\pi}{2}$ (x is a measure of angle AOM in radians), N is the foot of a perpendicular dropped from M onto OA, B is the point of intersection of the perpendicular to OA, erected from point A, with the extension of the segment OM. Then $MN = \sin x$, $ON = \cos x$, $AB = \tan x$. Since triangle OMA is contained in sector OMA which, in its turn, is contained in the triangle OBA, and the areas of the enumerated figures are equal, respectively, to $\frac{1}{2}\sin x$, $\frac{x}{2}$ and $\frac{1}{2}\tan x$, there hold inequalities $\sin x < x < \tan x$, $0 < x < \frac{\pi}{2}$. For the indicated values of x, $\sin x > 0$. Thus we have established the validity of the inequalities $0 < \sin x < x < \tan x$ $\left(\text{for } 0 < x < \frac{\pi}{2}\right)$.

Properties 1°, 2°, 3° can serve as the basis for defining the functions $\sin x$ and $\cos x$. It can be proved that *there exists a pair of functions defined for all the real values of the argument and satisfying the requirements* 1°, 2°, 3°, *and that this pair of functions is unique; the first of them we shall denote by* $\sin x$ *and the second by* $\cos x$.

The proof of this assertion is given in the supplement to the present chapter.

Make note of the fact that all the properties of trigonometric function known from the elementary course of mathematics can be derived from Properties 1°, 2°, and 3° of the functions $\sin x$ and $\cos x$.*

Let us prove the *continuity* of trigonometric functions at every point of the domain of their definition. We shall first establish the continuity of the function $y = \sin x$ at a point $x = 0$. Let $\{x_n\}$ be an arbitrary sequence of values of the argument x converging to point $x = 0$ on the right. From inequalities (4.7) we have $0 < \sin x_n < x_n$. By virtue of Theorem (3.14) it follows that the sequence $\{\sin x_n\}$ possesses the limit zero. Thus we have $\lim\limits_{x \to 0+0} \sin x = \sin 0 = 0$. Since inequalities $x < \sin x < 0$** are valid for $-\frac{\pi}{2} < x < 0$, a similar argument yields $\lim\limits_{x \to 0-0} \sin x = \sin 0$. We have established that at point $x = 0$ the function $y = \sin x$ is continuous on the right and on the left, i.e. it is continuous at that point. To prove the continuity of the function $y = \sin x$ at any point x of the infinite interval, we shall make use of the formula $\sin x'' - \sin x' = 2 \cos \frac{x''+x'}{2} \times \sin \frac{x''-x'}{2}$ which can be obtained from formulas (4.5). Let $\{x_n\}$ be a sequence of values of the argument converging to x. Putting $x'' = x_n$ and $x' = x$ in the last formula, we obtain

$$\lim_{n \to \infty} (\sin x_n - \sin x) = 2 \lim_{n \to \infty} \cos \frac{x_n + x}{2} \sin \frac{x_n - x}{2} = 0.$$

The validity of this inference follows from the fact that the sequence $\left\{\cos \frac{x_n+x}{2}\right\}$ is bounded*** and the sequence $\left\{\sin \frac{x_n-x}{2}\right\}$ is infinitesimal in accordance with the conclusions made above.

The *continuity of the function* $y = \cos x$ can be established in a similar way from the formula

$$\cos x'' - \cos x' = -2 \sin \frac{x''+x'}{2} \sin \frac{x''-x'}{2}.$$

The *continuity of the rest trigonometric functions* ($\tan x$, $\cot x$, $\sec x$, $\csc x$) at every point of the domain of their definition follows from Theorem 4.2.

---

* For instance, the equalities $\sin(-x) = -\sin x$, $\cos(-x) = \cos x$.

** These inequalities are obtained from inequalities (4.7) by substituting $-x$ for x and taking into account the formula $\sin(-x) = -\sin x$.

*** The third of formulas (4.5) enables us to infer that $|\cos x| \leqslant 1$ and $|\sin x| \leqslant 1$. It enables us to make a conclusion that the sequence $\left\{\cos \frac{x_n+x}{2}\right\}$ is bounded.

The domain of definition of every trigonometric function is divided into the intervals of monotonicity of that function*. The function $y = \sin x$ increases on every closed interval $\left[2k\pi - \frac{\pi}{2}, 2k\pi + \frac{\pi}{2}\right]$** and decreases on every interval $\left[(2k+1)\pi - \frac{\pi}{2}, (2k+1)\pi + \frac{\pi}{2}\right]$. The function $y = \cos x$ increases on every interval $[(2k-1)\pi, 2k\pi]$ and decreases on every interval $[2k\pi, (2k+1)\pi]$. The function $y = \tan x$ increases on every interval $\left(k\pi - \frac{\pi}{2}, k\pi + \frac{\pi}{2}\right)$. The function $y = \cot x$ decreases on every interval $((k-1)\pi, k\pi)$. The reader will easily establish the domains of increase and decrease of the functions $y = \sec x$ and $y = \csc x$.

Figures 4.22-4.27 depict the graphs of trigonometric functions.

4.5.7. Inverse trigonometric functions. The function $y = \arcsin x$ is defined as follows. Consider a function $y = \sin x$ on the closed interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. We have noted in the previous subsection that on this interval the function $y = \sin x$ increases, is continuous and has the interval $[-1, 1]$ as the set of its values. By virtue of the corollary of Lemma 1, there is a continuous, increasing inverse function of the function $y = \sin x$ on the closed interval $[-1, 1]$. We shall designate this function as $x = \arcsin y$. Replacing for that function the designation y of the argument by x and the designation x of the function by y, we shall obtain a function $y = \arcsin x$. Figure 4.28 shows the graph of this function.

The function $y = \arccos x$ can be defined by a complete analogy. Its domain of definition is the closed interval $[-1, 1]$ and the set of the values is the interval $[0, \pi]$. The indicated function decreases and is continuous on the interval $[-1, 1]$. Figure 4.29 illustrates the graph of the function $y = \arccos x$.

The functions $y = \arctan x$ and $y = \operatorname{arccot} x$ are defined as the inverse functions of the tangent and cotangent. These functions are defined, they are monotonic and continuous on the infinite interval. Figures 4.30 and 4.31 illustrate the graphs of these functions.

---

* The monotonicity of the functions $\sin x$ and $\cos x$ on the respective closed intervals can be easily established from the formulas

$$\sin x'' - \sin x' = 2 \cos \frac{x'' + x'}{2} \sin \frac{x'' - x'}{2}$$

and

$$\cos x'' - \cos x' = -2 \sin \frac{x'' + x'}{2} \sin \frac{x'' - x'}{2}.$$

** Here k is an arbitrary integer.

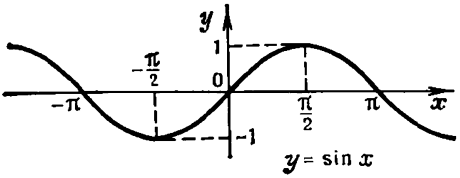


Fig. 4.22

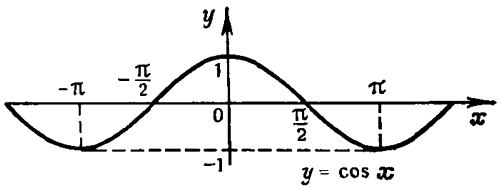


Fig. 4.23

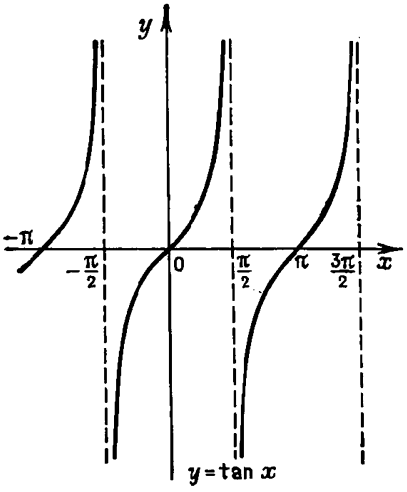


Fig. 4.24

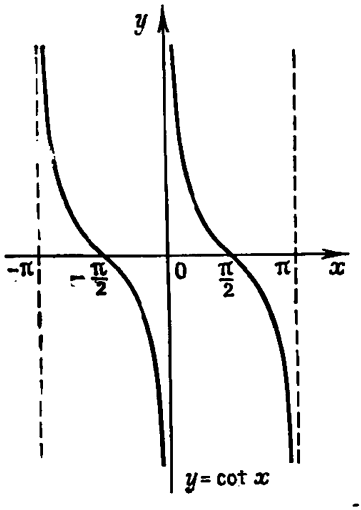


Fig. 4.25

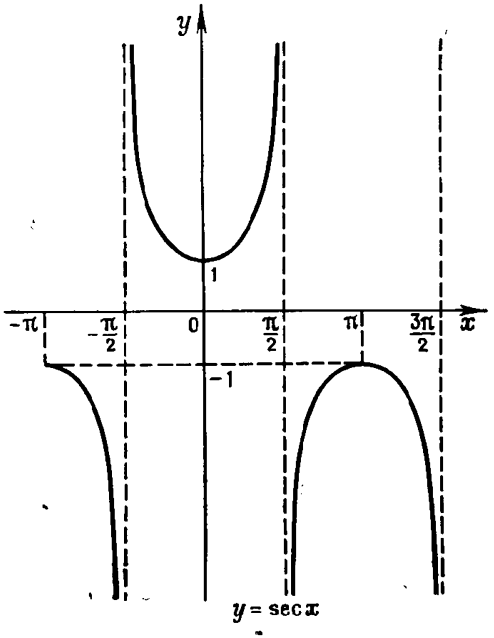


Fig. 4.26

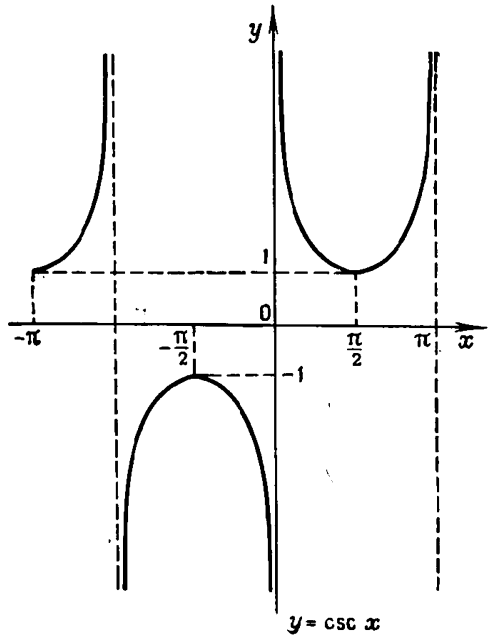


Fig. 4.27

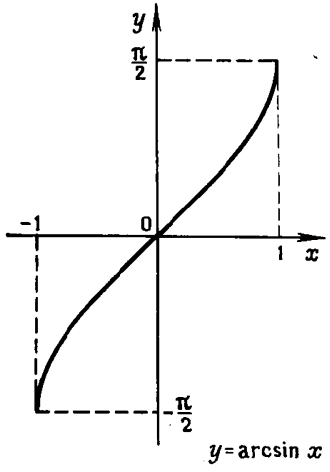


Fig. 4.28

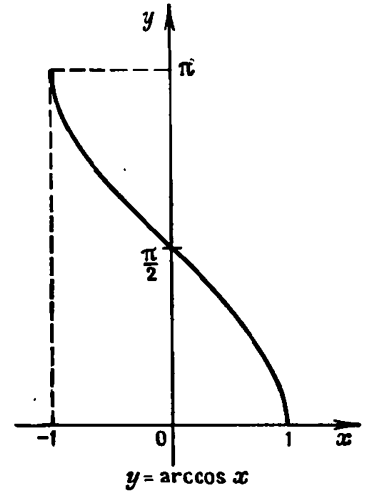


Fig. 4.29

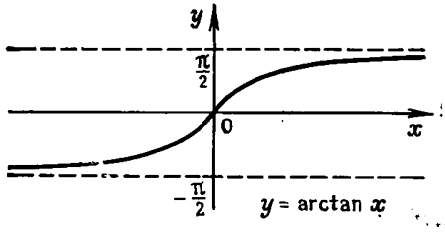


Fig. 4.30

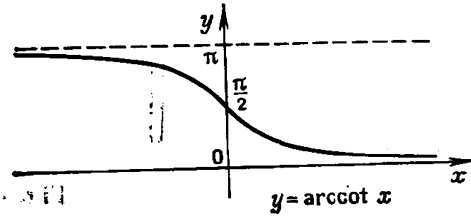


Fig. 4.31

4.6. LIMITS OF SOME FUNCTIONS

4.6.1. Preliminary remarks. It was pointed out in Chap. 1 that to calculate the derivatives of the functions $y = \sin x$ and $y = \log_a x$, it is necessary to prove the existence of the limiting values (or limits) of the function $\dfrac{\sin \frac{\Delta x}{2}}{\Delta x/2}$ for $\Delta x \to 0$ and of the function $\left(1 + \dfrac{\Delta x}{x}\right)^{\frac{x}{\Delta x}}$ for $\Delta x \to 0$ with a positive fixed x. This particular question is treated in the present subsection. We shall need a proposition concerning the limiting value of the function contained between two functions which have a common limit at a given point. This proposition is a functional analogue of Theorem 3.14.

Lemma 3. *Assume that functions $f(x)$, $g(x)$ and $h(x)$ are given in a certain δ-neighbourhood of a point a (with the exception, perhaps, of the point a itself), and let the functions $f(x)$ and $g(x)$ have the same limiting value at point a equal to b. If in the indicated neighbourhood of point a (except, perhaps, point a itself) there hold inequalities $f(x) \leqslant h(x) \leqslant g(x)$, then the limit of the function $h(x)$ at point a exists and is equal to b.*

Proof. Suppose $\{x_n\}$ is an arbitrary sequence of values of the argument x converging to a whose terms x_n belong to the indicated δ-neighbourhood of point a and are not equal to a, and assume that $\{f(x_n)\}$, $\{g(x_n)\}$, $\{h(x_n)\}$ are the corresponding sequences of values of the functions $f(x)$, $g(x)$ and $h(x)$. By the hypothesis,

$$\lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} g(x_n) = b \text{ and } f(x_n) \leqslant h(x_n) \leqslant g(x_n).$$

But then, by virtue of Theorem 3.14, $\lim\limits_{n\to\infty} h(x_n) = b$. Since $\{x_n\}$ is an arbitrary, converging to a, sequence of values of the argument, the last equality implies that $\lim\limits_{x\to a} h(x) = b$. The proof is complete.

4.6.2. The limit of the function $\dfrac{\sin x^*}{x}$ at point $x = 0$ (the first remarkable limit). Let us prove the following theorem.

Theorem 4.4. *The limit of the function $\dfrac{\sin x}{x}$ at point $x = 0$ exists and is equal to unity:*

$$\lim_{x\to 0} \frac{\sin x}{x} = 1. \tag{4.8}$$

* We have spoken of the function $\dfrac{\sin \frac{\Delta x}{2}}{\Delta x/2}$. If we denote $\dfrac{\Delta x}{2}$ by x, then we shall obtain the function $\dfrac{\sin x}{x}$. With this notation, the condition $\Delta x \to 0$ reduces to the condition $x \to 0$.

Proof. We have pointed out earlier that for $0 < x < \frac{\pi}{2}$ there hold inequalities $0 < \sin x < x < \tan x$ (see 4.5.6). Dividing these inequalities by $\sin x$ term-by-term, we obtain

$$1 < \frac{x}{\sin x} < \frac{1}{\cos x} \quad \text{or} \quad \cos x < \frac{\sin x}{x} < 1.$$

The last inequalities also hold true for the values of x satisfying the conditions $-\frac{\pi}{2} < x < 0$. To make sure of this fact, it is sufficient to note that $\cos x = \cos(-x)$ and $\frac{\sin x}{x} = \frac{\sin(-x)}{-x}$. Since $\cos x$ is a continuous function, we have $\lim\limits_{x \to 0} \cos x = 1$. Thus, all the conditions of Lemma 3 are satisfied for the functions $\cos x$, 1 and $\frac{\sin x}{x}$ in a certain δ-neighbourhood of point $x = 0$ (to be sure of this, use the notation $f(x) = \cos x$, $g(x) = 1$ and $h(x) = \frac{\sin x}{x}$ and put $\delta = \frac{\pi}{2}$). Consequently, $\lim\limits_{x \to 0} \frac{\sin x}{x} = \lim\limits_{x \to 0} \cos x = 1$. We have proved the theorem.

4.6.3. The limit of the function $\left(1 + \frac{1}{x}\right)^x$ for $x \to \infty$ (the second remarkable limit)*. Let us prove the following theorem.

Theorem 4.5. *The limit of the function $f(x) = \left(1 + \frac{1}{x}\right)^x$ exists for $x \to \infty$ and is equal to e:*

$$\lim_{x \to \infty} \left(1 + \frac{1}{x}\right)^x = e. \tag{4.9}$$

Proof. We have to prove that whatever the infinitely large sequence $\{x_k\}$ of the values of the argument of the function $f(x) = \left(1 + \frac{1}{x}\right)^x$, the corresponding sequence $\{f(x_k)\}$ of the values of that function has the number e as its limit. Let us consider the follow-

---

* The problem discussed earlier on the limit of the function $\left(1 + \frac{\Delta x}{x}\right)^{\frac{x}{\Delta x}}$ for Δx tending to zero and for the positive fixed x, reduces to the indicated problem. Indeed, if we put $\frac{\Delta x}{x} = \frac{1}{u}$, then, for $\Delta x \to 0$, $u \to \infty$ and $\left(1 + \frac{\Delta x}{x}\right)^{\frac{x}{\Delta x}} = \left(1 + \frac{1}{u}\right)^u$, and this function differs from the function $\left(1 + \frac{1}{x}\right)^x$ only by the designation of the argument.

ing four groups of infinitely large sequences of the values of the argument x.

1°. Infinitely large sequences $\{n_k\}$ whose terms are positive integers. This group includes, for instance, the sequence

$$2, 2, 1, 1, 3, 3, 2, 2, 4, 4, \ldots, n+1, n+1, n, n, \ldots$$

2°. Infinitely large sequences, whose terms, beginning with a certain number, consist of positive real numbers.

3°. Infinitely large sequences, whose terms, beginning with a certain number, consist of negative real numbers.

4°. Infinitely large sequences, containing infinitely many positive and negative real numbers*.

Note that quite an arbitrary infinitely large sequence of values of the argument belongs to one of the groups 2°, 3°, 4°. Therefore, the theorem will be proved if we carry out the proof for each of the groups 1°, 2°, 3°, 4°. (The sequence of group 1° is auxiliary.) Let $\{n_k\}$ be a sequence belonging to group 1°. We shall prove that $\lim\limits_{k\to\infty}\left(1+\frac{1}{n_k}\right)^{n_k}=e$. Let ε be any positive number. Since $\lim\limits_{n\to\infty}\left(1+\frac{1}{n}\right)^{n}=e$ (see 3.3.4.), there is a number N^* such that for $n\geqslant N^*$ there holds an inequality

$$\left|\left(1+\frac{1}{n}\right)^{n}-e\right|<\varepsilon.$$

Since the sequence $\{n_k\}$ is infinitely large and its terms are positive integers, there can be chosen a number N, for a positive N^*, such that the condition $n_k\geqslant N^*$ holds true when $k\geqslant N$. But, as has been indicated, for such integral n_k there holds an inequality

$$\left|\left(1+\frac{1}{n_k}\right)^{n_k}-e\right|<\varepsilon.$$

Consequently

$$\lim_{k\to\infty}\left(1+\frac{1}{n_k}\right)^{n_k}=e.$$

Now let us pass to the sequences of group 2°. Let $\{x_k\}$ be any sequence of the second group and N be the number beginning with which all the terms of that sequence are larger than unity. Assuming

---

* Since the function $\left(1+\frac{1}{x}\right)^{x}$ is not defined on the segment $[-1, 0]$ (because for the values of x belonging to that segment the expression $\left(1+\frac{1}{x}\right)$ is either negative or meaningless), it is natural to assume that the terms of the sequences 2°, 3°, 4° do not belong to the segment $[-1, 0]$.

$k \geqslant N$, we shall designate by n_k the *integral part* of x_k, $n_k = \{x_k\}$. Then

$$n_k \leqslant x_k < n_k + 1. \tag{4.10}$$

Note that $\{n_k\}$ and $\{n_k + 1\}$ are the sequences of the first group. From inequality (4.10) we have

$$\frac{1}{n_k+1} < \frac{1}{x_k} \leqslant \frac{1}{n_k}$$

or

$$1+\frac{1}{n_k+1} < 1+\frac{1}{x_k} \leqslant 1+\frac{1}{n_k}.$$

From these inequalities, using (4.10) once again, we obtain

$$\left(1+\frac{1}{n_k+1}\right)^{n_k} < \left(1+\frac{1}{x_k}\right)^{x_k} < \left(1+\frac{1}{n_k}\right)^{n_k+1}. \tag{4.11}$$

The limits of the sequences $\left\{\left(1+\frac{1}{n_k+1}\right)^{n_k}\right\}$ and $\left\{\left(1+\right.\right.$ $\left.\left.+\frac{1}{n_k}\right)^{n_k+1}\right\}$ are equal to e. In fact, the first of these sequences can be represented as the product of the sequences $\left\{\left(1+\right.\right.$ $\left.\left.+\frac{1}{n_k+1}\right)^{n_k+1}\right\}$ and $\left\{\left(1+\frac{1}{n_k+1}\right)^{-1}\right\}$, whose limits are equal, respectively*, to e and 1. The second sequence is the product of the sequences $\left\{\left(1+\frac{1}{n_k}\right)^{n_k}\right\}$ and $\left\{\left(1+\frac{1}{n_k}\right)\right\}$ whose limits are also equal to e and 1 respectively. By virtue of the inequalities (4.11) we have, in accordance with Theorem 3.14,

$$\lim_{k\to\infty}\left(1+\frac{1}{x_k}\right)^{x_k} = e.$$

Let us consider the sequences of the third group. If $\{x_k\}$ is an infinitely large sequence, whose terms, beginning with some number, are negative, then the sequence $\{z_k\}$, where $z_k = -1 - x_k$, is infinitely large and its terms, from some number onward, consist of positive real numbers. Therefore, $\{z_k\}$ is a sequence of the second group. Since

$$\left(1+\frac{1}{x_k}\right)^{x_k} = \left(1+\frac{1}{z_k}\right)^{z_k+1}$$

and

$$\lim_{k\to\infty}\left(1+\frac{1}{z_k}\right)^{z_k+1} = \lim_{k\to\infty}\left(1+\frac{1}{z_k}\right)^{z_k}\left(1+\frac{1}{z_k}\right) = e,$$

* In this case it is taken into account that $\{n_k\}$ belongs to group 1°.

we have

$$\lim_{k\to\infty}\left(1+\frac{1}{x_k}\right)^{x_k}=e.$$

To complete the proof, we must consider the sequences of the fourth group. Let $\{x_k\}$ be a sequence of this kind. We shall designate by $\{x_k'\}$ a subsequence of this sequence consisting of all the nonnegative terms* of the sequence $\{x_k\}$, and by $\{x_k''\}$ a subsequence consisting of all the negative terms of the sequence $\{x_k\}$**. Since, according to what has been proved,

$$\lim_{k\to\infty}\left(1+\frac{1}{x_k'}\right)^{x_k'}=e \text{ and } \lim_{k\to\infty}\left(1+\frac{1}{x_k''}\right)^{x_k''}=e$$

for any positive ε we can indicate a number N such that for $k\geqslant N$

$$\left|\left(1+\frac{1}{x_k'}\right)^{x_k'}-e\right|<\varepsilon \text{ and } \left|\left(1+\frac{1}{x_k''}\right)^{x_k''}-e\right|<\varepsilon,$$

i.e. for $k\geqslant N$

$$\left|\left(1+\frac{1}{x_k}\right)^{x_k}-e\right|<\varepsilon.$$

Consequently

$$\lim_{k\to\infty}\left(1+\frac{1}{x_k}\right)^{x_k}=e.$$

We have completed the proof of the theorem.

Remark. It follows from the theorem we have proved that

$$\lim_{x\to 0}(1+x)^{1/x}=e.$$

Indeed, let $\{x_n\}$ be any sequence, converging to zero, of values of the argument of the function $(1+x)^{1/x}$, whose terms x_n are nonzero. Then the sequence $\{z_n\}$, where $z_n=\frac{1}{x_n}$, is infinitely large (see Theorem 3.6). Since

$$(1+x_n)^{\frac{1}{x_n}}=\left(1+\frac{1}{z_n}\right)^{z_n} \text{ and } \lim_{n\to\infty}\left(1+\frac{1}{z_n}\right)^{z_n}=e,$$

we have

$$\lim_{n\to\infty}(1+x_n)^{1/x_n}=e,$$

and, consequently,

$$\lim_{x\to 0}(1+x)^{1/x}=e.$$

* Beginning with a certain number, these terms are strictly positive.

** As distinct from Chap. 3, we mark here the chosen subsequences with the signs ′ and ″, while the terms of the subsequence retain the numbers they had in the sequence $\{x_k\}$.

4.7. CONTINUITY AND LIMITING VALUES OF SOME COMPOSITE FUNCTIONS

4.7.1. Continuity and limiting values of some composite functions. We shall now prove the continuity of some composite functions.

1°. Let $x = \varphi(t)$ and $y = f(x)$ be basic elementary functions (see Sec. 5), the set $\{x\}$ of values of the function $x = \varphi(t)$ being the domain of definition of the function $y = f(x)$. It follows from the results obtained in 4.5 that the basic elementary functions are continuous at every point of the domain of definition. Therefore, in accordance with Theorem 4.5, the composite function $y = f[\varphi(t)]$, i.e. the superposition of two elementary functions, is continuous. For example, the function $y = \sin\frac{1}{x}$ is continuous at any point $x \neq 0$. To make sure of this fact, it is sufficient to consider the functions $x = t^{-1}$ and $y = \sin x$. The composite function $y = \sin t^{-1}$ differs from the function $y = \sin\frac{1}{x}$ only by the designation of the argument and, in accordance with what was said above, is continuous at any point $t \neq 0$. An analogous argument easily leads to the conclusion that the function $y = \ln \sin x$ is continuous at any point of each interval $(2k\pi,\ (2k+1)\pi)$*.

2°. Power-exponential expressions $u(x)^{v(x)}$.

Evidently, only the case $u(x) > 0$ is meaningful. It is easy to verify that if $u(x)$ and $v(x)$ are continuous at a point a and $u(x) > 0$ in the neighbourhood of point a, then the function $u(x)^{v(x)}$ is also continuous at point a.

Indeed, $u(x)^{v(x)} = e^{v(x)\ln u(x)}$. Since $\ln u(x)$ is a function continuous at point a, the function $v(x)\ln u(x)$ is also continuous at point a. But then the function $e^{v(x)\ln u(x)}$ is continuous at point a. Make note of the fact that the continuity property we have established makes it possible to assert that under the assumptions made $\lim\limits_{x\to a} u(x)^{v(x)} = u(a)^{v(a)}$.

3°. Limiting values of power-exponential expressions.

Let us elucidate the question concerning limiting values of power-exponential expressions $u(x)^{v(x)}$ in the case when $x \to a$. We shall assume that $u(x) > 0$, in a certain neighbourhood of point a.

It can be seen from the relation $u(x)^{v(x)} = e^{v(x)\ln u(x)}$ that the limit of the expression $u(x)^{v(x)}$, for $x \to a$, depends on the limit of the expression $v(x)\ln u(x)$.

1. Assume that $\lim\limits_{x\to a} v(x)\ln u(x) = b$.

Let us verify that in this case $\lim\limits_{x\to a} u(x)^{v(x)} = e^b$.

Indeed, the function

$$w(x) = \begin{cases} v(x)\ln u(x) & \text{for } x \neq a, \\ b & \text{for } x = a, \end{cases}$$

* In cases where $\sin x > 0$.

is continuous at point $x = a$. Therefore, the composite function $e^{w(x)}$ is continuous at that point. Consequently, $\lim\limits_{x \to a} e^{w(x)} = e^{w(a)} = e^b$. Since $\lim\limits_{x \to a} e^{w(x)} = \lim\limits_{x \to a} e^{v(x) \ln u(x)}$, $\lim\limits_{x \to a} u(x)^{v(x)}$ exists and is equal to e^b.

Making use of the information on the limiting values of e^w for $w \to -\infty$ and $w \to +\infty$ obtained in this chapter, it is easy to verify that

II. If $\lim\limits_{x \to a} v(x) \ln u(x) = -\infty$, then $\lim\limits_{x \to a} u(x)^{v(x)} = 0$.

III. If $\lim\limits_{x \to a} v(x) \ln u(x) = +\infty$, then $\lim\limits_{x \to a} u(x)^{v(x)} = +\infty$.

The relationship established between the limiting values of $u(x)^{v(x)}$ and $v(x) \ln u(x)$ makes it easy, in a number of cases, to find the limit of the function $u(x)^{v(x)}$ proceeding from the known values of the functions $u(x)$ and $v(x)$. Let us discuss, for example, the following cases.

(1) There exist $\lim\limits_{x \to a} u(x) > 0$ and $\lim\limits_{x \to a} v(x)$.

(2) $\lim\limits_{x \to a} u(x) = b$, $b > 1$, $\lim\limits_{x \to a} v(x) = +\infty$.

(3) $\lim\limits_{x \to a} u(x) = b$, $b > 1$, $\lim\limits_{x \to a} v(x) = -\infty$.

Let us verify that in case (1) $\lim\limits_{x \to a} u(x)^{v(x)} = [\lim\limits_{x \to a} u(x)]^{\lim\limits_{x \to a} v(x)}$. Indeed, since $\lim\limits_{x \to a} u(x) > 0$, by virtue of the continuity of a logarithmic function the limit $\lim\limits_{x \to a} \ln u(x)$ exists and is equal to $\ln [\lim\limits_{x \to a} u(x)]$. Therefore, there exists a limit $\lim\limits_{x \to a} v(x) \ln u(x) = \lim\limits_{x \to a} v(x) \ln [\lim\limits_{x \to a} u(x)]$.

It follows, in accordance with I, that

$$[\lim_{x \to a} u(x)^{v(x)} = e^{\lim\limits_{x \to a} v(x) \ln u(x)} = e^{\lim\limits_{x \to a} v(x) \ln [\lim\limits_{x \to a} u(x)]} = [\lim_{x \to a} u(x)]^{\lim\limits_{x \to a} v(x)}.$$

In case (2) $\lim\limits_{x \to a} v(x) \ln u(x) = +\infty$, and therefore, in accordance with III, $\lim\limits_{x \to a} u(x)^{v(x)} = +\infty$.

In case (3) $\lim\limits_{x \to a} v(x) \ln u(x) = -\infty$, and therefore, in accordance with II, $\lim\limits_{x \to a} u(x)^{v(x)} = 0$.

In conclusion we shall mention three cases when finding the limiting value of $u(x)^{v(x)}$ requires additional calculations.

1. *Indeterminate form* 1^∞:

$$\lim_{x \to a} u(x) = 1, \qquad \lim_{x \to a} v(x) = \infty.$$

2. *Indeterminate form* 0^0:

$$\lim_{x \to a} u(x) = 0, \qquad \lim_{x \to a} v(x) = 0.$$

3. *Indeterminate form* ∞^0:

$$\lim_{x \to a} u(x) = +\infty, \qquad \lim_{x \to a} v(x) = 0.$$

For the first of these cases we shall cite a formula convenient for applications.

We shall transform the expression $u(x)^{v(x)}$ as follows:

$$u(x)^{v(x)} = \left\{[1+(u(x)-1)]^{\frac{1}{u(x)-1}}\right\}^{[u(x)-1]v(x)}.$$

Next we shall put

$$U(x) = [1+(u(x)-1)]^{\frac{1}{u(x)-1}} \quad \text{and} \quad V(x) = [u(x)-1]\,v(x)$$

so that

$$u(x)^{v(x)} = U(x)^{V(x)}.$$

Since $\lim\limits_{x\to a} U(x) = e$ (see the remark to Theorem 4.5) and $e > 1$, the value of $\lim\limits_{x\to a} u(x)^{v(x)} = \lim\limits_{x\to a} U(x)^{V(x)}$ *depends* on the limit of the function $V(x)$ at point a, i.e. on $\lim\limits_{x\to a}[u(x)-1]\,v(x)$, namely,

if $\lim\limits_{x\to a}[u(x)-1]\,v(x) = \lim\limits_{x\to a} V(x) = c$, then $\lim\limits_{x\to a} u(x)^{v(x)} = \lim\limits_{x\to a} U(x)^{V(x)} = e^c$ (see case (1)):

if $\lim\limits_{x\to a}[u(x)-1]\,v(x) = +\infty$, then $\lim\limits_{x\to a} u(x)^{v(x)} = +\infty$ (see case (2));

if $\lim\limits_{x\to a}[u(x)-1]\,v(x) = -\infty$, then $\lim\limits_{x\to a} u(x)^{v(x)} = 0$ (see case (3)).

Thus we obtain the following formula:

$$\lim_{x\to a} u(x)^{v(x)} = e^{\lim\limits_{x\to a}[u(x)-1]\,v(x)}$$

Indeterminate forms 2 and 3 can be reduced to indeterminate form 1 in the following way.

We assume that

$$U(x) = e^{v(x)}, \quad V(x) = \ln u(x).$$

It is evident that $\lim\limits_{x\to a} U(x) = 1$ and $\lim\limits_{x\to a} V(x) = \pm\infty$. Besides,

$$u(x)^{v(x)} = [e^{V(x)}]^{\ln U(x)} = e^{\ln U(x)^{V(x)}} = U(x)^{V(x)}$$

Example. It is required to find $\lim\limits_{x\to 0}[\cos x]^{\frac{1}{\sin^2 x}}$. Since $\lim\limits_{x\to 0}\cos x = 1$ and $\lim\limits_{x\to 0}\frac{1}{\sin^2 x} = \infty$, we have indeterminate form 1^∞. We make use of the formula $\lim\limits_{x\to 0} u(x)^{v(x)} = e^{\lim\limits_{x\to 0}[u(x)-1]v(x)}$ obtained above. Then we have

$$\lim_{x\to 0}[u(x)-1]\,v(x) = \lim_{x\to 0}[\cos x - 1]\frac{1}{\sin^2 x} =$$

$$= \lim_{x\to 0}\left[-2\sin^2\frac{x}{2}\right]\frac{1}{4\sin^2\frac{x}{2}\cos^2\frac{x}{2}} = -\frac{1}{2}\lim_{x\to 0}\frac{1}{\cos^2\frac{x}{2}} = -\frac{1}{2}$$

Therefore

$$\lim_{x\to 0}[\cos x]^{\frac{1}{\sin^2 x}}=e^{-\frac{1}{2}}=\frac{1}{\sqrt{e}}.$$

4°. Limiting values of some composite functions. Let us prove the validity of the following equalities:

$$\left.\begin{aligned}&\lim_{x\to 0}\frac{\sqrt[n]{1+x}-1}{x}=\frac{1}{n},\quad \lim_{x\to 0}\frac{\ln(1+x)}{x}=1,\\&\lim_{x\to 0}\frac{e^x-1}{x}=1,\quad \lim_{x\to 0}\frac{1-\cos x}{x^2}=\frac{1}{2}.\end{aligned}\right\}\tag{4.12}$$

(1) Let us consider the first of these limits. We have

$$\frac{\sqrt[n]{1+x}-1}{x}=\frac{(1+x)^{1/n}-1}{x}=$$

$$=\frac{[(1+x)^{1/n}-1]\left[(1+x)^{\frac{n-1}{n}}+(1+x)^{\frac{n-2}{n}}+\ldots+(1+x)^{1/n}+1\right]}{x\left[(1+x)^{\frac{n-1}{n}}+(1+x)^{\frac{n-2}{n}}+\ldots+(1+x)^{1/n}+1\right]}=$$

$$=\frac{[(1+x)^{1/n}]^n-1}{x\left[(1+x)^{\frac{n-1}{n}}+(1+x)^{\frac{n-2}{n}}+\ldots+(1+x)^{1/n}+1\right]}=$$

$$=\frac{1}{(1+x)^{\frac{n-1}{n}}+(1+x)^{\frac{n-2}{n}}+\ldots+(1+x)^{1/n}+1}.$$

Since for $x\to 0$ the denominator of the last expression has a limit equal to n (the function $(1+x)^{k/n}$ is continuous at point $x=0$ and, therefore, $\lim\limits_{x\to 0}(1+x)^{k/n}=1$), it follows that $\lim\limits_{x\to 0}\dfrac{\sqrt[n]{1+x}-1}{x}=\dfrac{1}{n}$.

(2) Let us now prove the second equality (4.12). We have $\dfrac{\ln(1+x)}{x}=\ln(1+x)^{1/x}$. Let us make the definition of the function $f(x)=(1+x)^{1/x}$ complete putting $f(0)=\lim\limits_{x\to 0}f(x)=\lim\limits_{x\to 0}(1+x)^{1/x}=e$. As a result we shall obtain the function $f(x)$ continuous at point $x=0$. Then the function $\ln f(x)$ will also be continuous at point zero and, therefore, $\lim\limits_{x\to 0}\ln(1+x)^{1/x}=\ln f(0)=\ln e=1$. Thus we have $\lim\limits_{x\to 0}\dfrac{\ln(1+x)}{x}=1$.

(3) We shall now prove the validity of the third equality (4.12). We shall put $x=\ln(1+u)$ and make note of the fact

that for $x \to 0$ the variable u tends to zero. We have $\frac{e^x-1}{x} = \frac{u}{\ln(1+u)}$, whence it follows that

$$\lim_{x\to 0}\frac{e^x-1}{x} = \lim_{u\to 0}\frac{u}{\ln(1+u)} = 1.$$

(4) Let us prove the validity of the last equality (4.12). We have

$$\frac{1-\cos x}{x^2} = \frac{2\sin^2\frac{x}{2}}{x^2} = \frac{1}{2}\,\frac{\sin^2\frac{x}{2}}{(x/2)^2}.$$

Since $\lim\limits_{x\to 0}\frac{\sin^2\frac{x}{2}}{(x/2)^2} = 1$ (see 4.8)), it follows that $\lim\limits_{x\to 0}\frac{1-\cos x}{x^2} = \frac{1}{2}$.

Making use of relations (4.8), (4.12), equality (4.1) and the symbol $o(x)$ (see 4.2.3), it is easy to ascertain the validity of the following formulas:

$$\left.\begin{aligned} &\sin x = x + o(x),\\ &\sqrt[n]{1+x} = 1 + \frac{x}{n} + o(x),\\ &\ln(1+x) = x + o(x),\\ &e^x = 1 + x + o(x),\\ &\cos x = 1 - \frac{x^2}{2} + o(x^2). \end{aligned}\right\} \qquad (4.13)$$

By way of example we shall prove the validity of the first formula. Since $\lim\limits_{x\to 0}\frac{\sin x}{x} = 1$, we have, by virtue of (4.1), $\frac{\sin x}{x} = 1 + \alpha(x)$, where $\alpha(x)$ is an infinitesimal function at point $x = 0$. It follows from the last formula that $\sin x = x + x\alpha(x)$. Inasmuch as $x\alpha(x) = o(x)$, we have $\sin x = x + o(x)$.

4.7.2. The concept of an elementary function. The class of elementary functions. An important role in applications is played by a class of functions obtained from basic elementary functions by means of a finite number of arithmetical operations as well as by means of superposition of these functions. The functions $x^3 + 3\cos 2x$, $\ln[\sin 3x] - e^{\arctan\sqrt{x}}$, for instance, belong to this class. This class of functions will be called the *class of elementary functions* and every function belonging to this class will be called an *elementary function.*

It might be well to point out the following property of elementary functions: *they are continuous at every point of the domain of definition**.

This property follows immediately from Theorems 4.2 and 4.3 and from the continuity of basic elementary functions at every point of the domain of definition.

4.8. CLASSIFICATION OF POINTS OF DISCONTINUITY OF A FUNCTION

4.8.1. Points of discontinuity of a function and their classification. In 4.3.1 we defined points of discontinuity of a function as the points at which the function does not possess the property of continuity. We shall also use this term, the points of discontinuity of a function, for the points at which the function is not defined but whose any ε-neighbourhood possesses points of the domain of definition of the function.

Let us discuss possible types of points of discontinuity of a function.

1°. *Removable discontinuity. Point a is said to be a point of removable discontinuity of the function $y = f(x)$ if the limiting value of that function at point a exists but the function $f(x)$ is either not defined at that point or its particular value $f(a)$ at point a is not equal to the limiting value.*

For example, the function

$$f(x) = \begin{cases} \dfrac{\sin x}{x} & \text{for} \quad x \neq 0, \\ 2 & \text{for} \quad x = 0 \end{cases}$$

has a removable discontinuity at point zero because the limit of this function at point $x = 0$ is equal to unity and its particular value is 2. If the function $f(x)$ has a discontinuity of the indicated kind at point a, this discontinuity *can be removed* without changing the values of the function at points differing from a. For that purpose it is sufficient to assume that the value of the function at point a is equal to its limiting value at that point. Thus, if we put $f(0) = 1$ in the example being discussed, then $\lim\limits_{x \to 0} f(x) = f(0)$ and the function will be continuous at point $x = 0$.

Remark. Points of removable discontinuity are encountered in practice in the cases of concentrated distributions of physical quantities.

---

* If the domain of definition of the function turns out to consist of separate isolated points, then it is natural to assume that by definition the function is continuous at each of those points.

2°. *Discontinuity of the 1st kind. Point a is called a point of discontinuity of the 1st kind if at that point the function f (x) has finite but not equal to each other right-hand and left-hand limiting values:*

$$\lim_{x\to a+0} f(x) \neq \lim_{x\to a-0} f(x) \quad (\text{or } f(a+0) \neq f(a-0)).$$

1. For the function $f(x) = \operatorname{sgn} x$, point $x = 0$ is a point of discontinuity of the 1st kind (see Fig. 4.4). Indeed, since

$$\operatorname{sgn} x = \begin{cases} 1 & \text{for } x > 0, \\ 0 & \text{for } x = 0, \\ -1 & \text{for } x < 0, \end{cases}$$

it follows that

$$\lim_{x\to 0+0} \operatorname{sgn} x = 1, \quad \lim_{x\to 0-0} \operatorname{sgn} x = -1.$$

2. The function $f(x) = \frac{1}{1+2^{1/x}}$ defined everywhere except point $x = 0$ has at that point a discontinuity of the 1st kind (Fig. 4.32). In fact, if $\{x_n\}$ is a sequence, converging to zero, whose

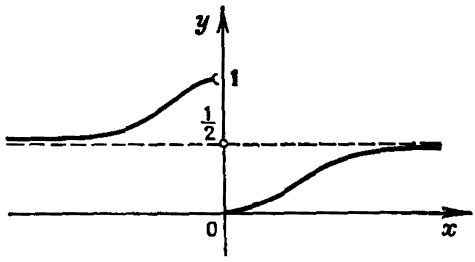


Fig. 4.32

terms are positive, then $\left\{\frac{1}{x_n}\right\}$ is an infinitely large sequence with positive terms and, therefore, $\{1+2^{1/x_n}\}$ is also an infinitely large sequence. But in that case the sequence $\left\{\frac{1}{1+2^{1/x_n}}\right\}$ is infinitesimal and, therefore, $\lim\limits_{x\to 0+0} f(x) = 0$. Now if $\{x_n\}$ is a sequence, converging to zero, whose terms are negative, then $\left\{\frac{1}{x_n}\right\}$ is an infinitely large sequence with negative terms and, therefore, $\lim\limits_{n\to\infty} 2^{1/x_n} = 0$. Consequently, $\lim\limits_{x\to 0-0} f(x) = 1$.

3°. *Discontinuity of the 2nd kind. Point a is called a point of discontinuity of the second kind if at that point the function f (x) does not*

possess at least one of its one-sided limits, or if at least one of its one-sided limits is infinite.

Let us discuss, for instance, a function $f(x) = \sin \frac{1}{x}$ (Fig. 4.33)*. This function has neither the right-hand nor the left-hand limit at

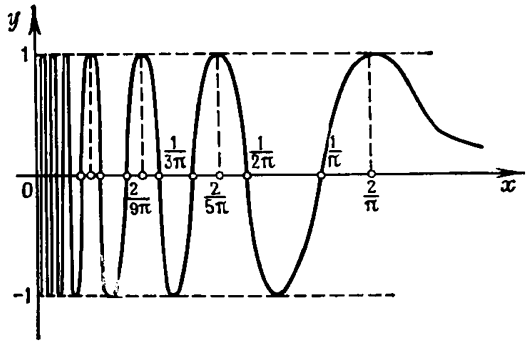


Fig. 4.33

point $x = 0$. Indeed, let us consider the following sequences of values of the argument converging to zero from the right:

$$\frac{2}{\pi}, \frac{2}{5\pi}, \frac{2}{9\pi}, \ldots, \frac{2}{(4n-3)\pi}, \ldots$$

and

$$\frac{1}{\pi}, \frac{1}{2\pi}, \frac{1}{3\pi}, \ldots, \frac{1}{n\pi}, \ldots$$

The corresponding sequences of values of the function $y = \sin(1/x)$ have the following form:

$$1, 1, 1, \ldots, 1, \ldots$$

and

$$0, 0, 0, \ldots, 0, \ldots$$

The first sequence has the limit unity and the second has the limit zero. Consequently, the function $f(x) = \sin(1/x)$ has no right-hand limit at point $x = 0$. Now since $\sin(1/-x) = -\sin(1/x)$, the function has no left-hand limit at that point either.

Another function which can be cited as an example of a function having points of discontinuity of the second kind is $y = \cot x$ (see Fig. 4.25). This function has a discontinuity of the 2nd kind at each of the points πn, $n = 0, \pm 1, \pm 2, \ldots$

* Figure 4.33 is purely illustrative.

4.8.2. Piecewise-continuous functions. The function $y = f(x)$ is said to be *piecewise-continuous* on the closed interval $[a, b]$ if it is continuous at all the interior points $[a, b]$ except, perhaps, a finite number of points at which it has a discontinuity of the first kind, and, besides, if it has one-sided limiting values at points a and b. A function is called piecewise-continuous on an interval or an infinite number line if it is piecewise-continuous on any closed interval belonging to them. For example, the function $f(x) = [x]$* is piecewise-continuous on any closed interval as well as on any number line.

SUPPLEMENT TO CHAPTER 4

PROOF OF THE ASSERTION MADE IN 4.5.6

This supplement contains the proof of the assertion made in 4.5.6. For the sake of convenience we shall formulate the assertion here as follows.

There exists a unique pair of functions $S(x)$ and $C(x)$ defined throughout the number line and satisfying the following three requirements:

1°. *For any real numbers x', x'' and x there hold relations*

$$S(x'+x'') = S(x')\,C(x'') + C(x')\,S(x''),$$

$$C(x'+x'') = C(x')\,C(x'') - S(x')\,S(x''),$$

$$S^2(x) + C^2(x) = 1. \qquad (4.5)^{**}$$

2°.

$$\left.\begin{aligned} &S(0) = 0, \quad C(0) = 1, \\ &S\left(\frac{\pi}{2}\right) = 1, \quad C\left(\frac{\pi}{2}\right) = 0. \end{aligned}\right\} \qquad (4.6)$$

3°. *For $0 < x < \frac{\pi}{2}$ there hold inequalities*

$$0 < S(x) < x. \qquad (4.7)$$

We shall divide the proof of the assertion into two parts, namely, first we shall prove the *uniqueness* and then the *existence* of the functions $S(x)$ and $C(x)$ satisfying the requirements 1°, 2° and 3°.

1. The proof of the uniqueness. To prove the uniqueness, it is sufficient to ascertain the following two assertions:

(1) *The functions $S(x)$ and $C(x)$ possessing the indicated properties are continuous on the entire number line.*

(2) *The values of the functions $S(x)$ and $C(x)$ are uniquely defined on a certain, dense everywhere, set of points of the number line****.

Indeed, by virtue of the continuity of the functions $S(x)$ and $C(x)$, their particular values at each point x of the number line are equal to their limiting

* Recall that the symbol $[x]$ denotes the integral part of the number x.

** The numbering of the formulas is the same as in 4.5.6. Note that in these formulas we have substituted $S(x)$ and $C(x)$ for the designations of the functions $\sin x$ and $\cos x$ respectively.

*** The set $\{x\}$ of points of the number line is said to be dense everywhere on that line if there are infinitely many points of the set $\{x\}$ in any ε-neighbourhood of every point of that line.

values at that point. If we now consider the sequence, converging to x, of the values of the argument whose terms belong to the earlier mentioned set of points dense everywhere, then, in accordance with assertion 2 formulated above, the corresponding sequences of values of the functions $S(x)$ and $C(x)$ are determined uniquely and, therefore, the limits of those sequences are also determined uniquely. But these limits are precisely the particular values of the functions $S(x)$ and $C(x)$ at point x. Consequently, the functions $S(x)$ and $C(x)$ are determined uniquely on the entire number line.

(1) Before passing to the proof of the continuity of the functions $S(x)$ and $C(x)$ we shall establish certain formulas.

Putting $x' = x$, $x'' = -x$ in the first two relations (4.5) and taking into account that $S(0) = 0$, $C(0) = 1$, we obtain

$$\left.\begin{aligned} 0 &= S(x)\cdot C(-x) + C(x)\cdot S(-x), \\ 1 &= C(x)\cdot C(-x) - S(x)\cdot S(-x). \end{aligned}\right\} \tag{4.14}$$

Let us multiply relations (4.14) by $S(x)$ and $C(x)$ respectively and sum up the resulting equations. Taking into account that $S^2(x) + C^2(x) = 1$, we obtain $C(-x) = C(x)$.

By a complete analogy, multiplying relations (4.14) by $C(x)$ and $-S(x)$ respectively and summing them up, we shall obtain $S(-x) = -S(x)$. Thus we see that $C(x)$ is an *even* function and $S(x)$ is an *odd* function*.

But then, using the first of formulas (4.5), we obtain

$$S(x'') = S\left(\frac{x'+x''}{2} + \frac{x''-x'}{2}\right) = S\left(\frac{x'+x''}{2}\right) C\left(\frac{x''-x'}{2}\right) +$$

$$+ C\left(\frac{x'+x''}{2}\right) S\left(\frac{x''-x'}{2}\right)$$

and similarly

$$S(x') = S\left(\frac{x'+x''}{2} + \frac{x'-x''}{2}\right) = S\left(\frac{x'+x''}{2}\right) C\left(-\frac{x''-x'}{2}\right) +$$

$$+ C\left(\frac{x'+x''}{2}\right) S\left(-\frac{x''-x'}{2}\right) = S\left(\frac{x'+x''}{2}\right) C\left(\frac{x''-x'}{1}\right) -$$

$$- C\left(\frac{x'+x''}{2}\right) S\left(\frac{x''-x'}{2}\right).$$

Subtracting the last two formulas term-by-term, we obtain

$$S(x'') - S(x') = 2C\left(\frac{x'+x''}{2}\right) S\left(\frac{x''-x'}{2}\right). \tag{4.15}$$

Let us now prove the continuity of the functions $C(x)$ and $S(x)$ at any point x of the number line. Note the fact that the continuity of the function $S(x)$ *at point* $x = 0$ *on the right* follows immediately from relation (4.7) and from the equality $S(0) = 0$. Indeed, if $\{x_n\}$ is an arbitrary sequence of values of the argument converging to zero on the right, then we can infer from the relation $0 < S(x_n) < x_n$ that the corresponding sequence of values of the function $\{S(x_n)\}$ converges to zero, i.e. to the particular value $S(0)$.

The continuity of the function $S(x)$ *at point* $x = 0$ *on the left* follows from the *oddness* of that function. Thus, the function $S(x)$ is *continuous at point* $x = 0$.

---

* The function $f(x)$ is said to be *odd* if $f(-x) = -f(x)$ and *even* if $f(-x) = f(x)$.

The continuity of $S(x)$ *at any point* x *follows* from relation (4.15). In fact, suppose x *is any point of the infinite interval,* $\{x_n\}$ *is an arbitrary sequence of values of the argument converging to* x. Putting $x' = x$, $x'' = x_n$ in (4.15), we shall have

$$S(x_n) - S(x) = 2C\left(\frac{x+x_n}{2}\right) S\left(\frac{x_n - x}{2}\right). \tag{4.16}$$

Since $S(x)$ is continuous at point zero and $S(0)=0$, we obtain $\lim\limits_{n\to\infty} S \times \times\left(\frac{x_n - x}{2}\right) = 0$. Inasmuch as the sequence $\left\{C\left(\frac{x+x_n}{2}\right)\right\}$ is *bounded**, the right-hand side (and, consequently, the left-hand side) of (4.16) possesses the limit zero. But this means that $\lim\limits_{n\to\infty} S(x_n) = S(x)$, i.e. that the function $S(x)$ is continuous at point x.

The continuity of the function $C(x)$ can be proved by analogy, with the only difference that instead of (4.15) we must obtain the formula

$$C(x'') - C(x') = -2S\left(\frac{x''+x'}{2}\right) S\left(\frac{x''-x'}{2}\right).$$

(2) Let us now prove that the values of the functions $S(x)$ and $C(x)$ are determined uniquely at points $\frac{p\pi}{2^n}$, where p is a positive or negative integer and n is a positive integer. It might be well to point out that such points form everywhere a dense set of points of the number scale. We shall first of all establish some properties of the functions $S(x)$ and $C(x)$. First we shall ascertain that these functions are *periodic with period* 2π**. Indeed, assuming $x'' = x + 2\pi$ and $x' = x$ in (4.15), we obtain

$$S(x + 2\pi) - S(x) = 2C(x + \pi) S(\pi).$$

Since $S(\pi) = S\left(\frac{\pi}{2} + \frac{\pi}{2}\right) = 2S\left(\frac{\pi}{2}\right) C\left(\frac{\pi}{2}\right) = 0$, it follows, from the last relation, that

$$S(x + 2\pi) = S(x),$$

i.e. the function $S(x)$ is periodic with period 2π. This implies, in particular, that $S(2\pi) = 0$.

Putting in the second formula (4.5) $x' = x$ and $x'' = 2\pi$ and taking into account that $S(2\pi) = 0$, we shall find

$$C(x + 2\pi) = C(x) C(2\pi).$$

Since $C(2\pi) = 1$ (and this can be easily ascertained by applying formulas (4.5) first for $x' = \pi/2$ and $x'' = \pi/2$ and then for $x' = \pi$ and $x'' = \pi$) we have

$$C(x + 2\pi) = C(x).$$

Thus the periodicity of $C(x)$ has been established as well.

The periodicity of the functions $S(x)$ and $C(x)$ enables us to restrict our discussion to the interval $[0, 2\pi]$. We shall now establish the signs of the values of the functions $S(x)$ and $C(x)$ at various points of that interval. It follows from

* It follows from the relation $S^2(x) + C^2(x) = 1$ that $|C(x)| \leqslant 1$ for all x, and this implies the boundedness of the sequence $\left\{C\left(\frac{x+x_n}{2}\right)\right\}$.

** The function $f(x)$ is said to be *periodic with period* $a > 0$ if, for any x, there holds relation $f(x+a) = f(x)$.

(4.6), (4.7) and the continuity of $S(x)$ that in the interval $[0, \pi/2]$ the values of the function $S(x)$ are nonnegative, and that in the indicated interval the function $S(x)$ vanishes only at point $x = 0$. Since $S(\pi - x) = S(\pi) C(-x) - -C(\pi) S(x)$* and $S(\pi) = 0$, $C(\pi) = -1$, we have $S(\pi - x) = S(x)$. Therefore, on the interval $[\pi/2, \pi]$ the values of the function $S(x)$ are nonnegative, with the function $S(x)$ vanishing on that interval only at point $x = \pi$. The formula $S(2\pi - x) = -S(x)$, which can be obtained in the same way as the formula $S(\pi - x) = S(x)$, implies that on the interval $[\pi, 2\pi]$ the values of the function $S(x)$ are nonpositive, with the functions $S(x)$ vanishing only at the end points of that interval. Arguing in a similar fashion we can make sure that the function $C(x)$ is nonnegative on the intervals $[0, \pi/2]$ and $[3\pi/2, 2\pi]$ and nonpositive on the interval $[\pi/2, 3\pi/2]$ and vanishes only at points $\pi/2$ and $3\pi/2$.

To complete the proof of the uniqueness of the functions $S(x)$ and $C(x)$, we shall need some formulas which we shall now derive. Note, first of all, that (4.5) yields the following formulas**:

$$S^2\left(\frac{x}{2}\right) = \frac{1 - C(x)}{2}, \quad C^2\left(\frac{x}{2}\right) = \frac{1 + C(x)}{2}. \tag{4.17}$$

Assuming in these formulas $x = x' + x''$ and applying formulas (4.5) once again, we obtain the formulas required:

$$S^2\left(\frac{x' + x''}{2}\right) = \frac{1 - C(x')\,C(x'') + S(x')\,S(x'')}{2},$$

$$C^2\left(\frac{x' + x''}{2}\right) = \frac{1 + C(x')\,C(x'') - S(x')\,S(x'')}{2}.$$

These formulas show that if the values of the functions $S(x)$ and $C(x)$ at points x' and x'' are known, then the values of these functions at point $\frac{x' + x''}{2}$ are determined uniquely, for it follows from the above argument that the signs of the functions $S(x)$ and $C(x)$ are known at each point of the closed interval $[0, 2\pi]$ and, consequently, by virtue of their periodicity with period 2π, at any point x of the number line as well. Proceeding from the known and uniquely determined values of $S(x)$ and $C(x)$ at points $0, \pi/2, \pi, 2\pi$ of the interval $[0, 2\pi]$, we can compute uniquely the values of those functions at all the points of the form $p\pi/2^n$ of the interval $[0, 2\pi]$ by the successive application of the formulas just obtained (p and n are nonnegative integers with $p \leqslant 2^{n+1}$). Since the set of points $p\pi/2^n$ is dense everywhere on the interval $[0, 2\pi]$ then, according to what was said at the beginning of the proof of uniqueness, the functions $S(x)$ and $C(x)$ are determined uniquely on the whole number line.

2. The proof of the existence. We shall now prove a more general assertion.

There exist functions $S(x)$ and $C(x)$, defined and continuous on the entire number line, which satisfy the following requirements:

1°. *For any real numbers x', x'' and x there hold relations*

$$\left.\begin{aligned} S(x' + x'') &= S(x')\,C(x'') + C(x')\,S(x''), \\ C(x' + x'') &= C(x')\,C(x'') - S(x')\,S(x''), \\ S^2(x) + C^2(x) &= 1. \end{aligned}\right\} \tag{4.5^1}$$

---

* This formula follows from the first formula (4.5) and the oddness of the function $S(x)$.

** It is sufficient to substitute in the second formula (4.5) $x' = x'' = x/2$ and in the third formula $x/2$ for x.

2°.

$$\left.\begin{aligned} S(0)=0, \quad C(0)=1, \\ S(d)=1, \quad C(d)=0, \end{aligned}\right\} \tag{4.6^1}$$

where d is a certain preassigned number.

3°. *There exists a positive number L such that for $0 < x < d$ there hold inequalities*

$$0 < S(x) < Lx, \tag{4.7^1}$$

with $L = 1$ when $d = \pi/2$.

Proof. *Let us first determine the values of the functions $S(x)$ and $C(x)$ on the set $\{s\}$ of the points of the interval $[0, d]$ each of which can be represented as $s = pd/2^n$, where p and n are nonnegative integers, with $p < 2^n$.* We shall begin with determining the values of those functions at points $s_n = d/2^n$, $n = 0, 1, \ldots$. Since $s_{n+1} = s_n/2$, we can use formulas (4.17) and put

$$S(s_{n+1}) = S\left(\frac{s_n}{2}\right) = +\sqrt{\frac{1-C(s_n)}{2}}, \quad C(s_{n+1}) = \overline{C}\left(\frac{s_n}{2}\right) = +\sqrt{\frac{1+C(s_n)}{2}}. \tag{4.18}$$

The values of the functions $S(x)$ and $C(x)$ are determined at all the points $s_n = d/2^n$ from $C(d) = C(d/2^0) = C(s_0) = 0$ with the aid of recurrence formulas (4.18). In addition to the indicated values of $S(x)$ and $C(x)$ at points s_n, we shall determine the values of these functions at points 0 and d in a way shown in (4.6^1).

Now we shall pass to determining the values of $S(x)$ and $C(x)$ at all the points of the set $\{s\}$, $s = pd/2^n$, where p and n are nonnegative integers, $p < 2^n$. It is known that any integral positive number can be uniquely represented as the sum of the integral powers of the number 2*:

$$p = \sum_{i=1}^{n} a_i 2^{n-i}{}^{**}$$

where a_i is equal either to zero or to unity. Therefore,

$$s = \frac{pd}{2^n} = \sum_{i=1}^{n} \frac{a_i d}{2^i} = \sum_{i=1}^{n} a_i s_i. \tag{4.19}$$

Thus, every value of s can be represented as the finite sum of the numbers s_i for each of which the values of $S(s_i)$ and $C(s_i)$ were determined above. We can now find the values of $S(x)$ and $C(x)$ at the points of the set $\{s\}$ by making use of formulas (4.5^1). But we must be sure that the successive application of these formulas leads to the same result irrespective of the way the summands s_i are grouped in formula (4.19).

For example, we can put $s = x' + x''$, where $x' = a_1 s_1$ and $x'' = \sum_{i=2}^{n} a_i s_i$, and then compute $S(s)$ by means of the first formula (4.5^1). But we can also put $x' = a_1 s_1 + a_2 s_2$ and $x'' = \sum_{i=3}^{n} a_i s_i$. To make sure that the successive application of formulas (4.5^1) will lead to the same result irrespective of the way of

* In the binary system of notation the integral number p is uniquely represented as a symbol consisting of zeros and unities. That symbol is the abbreviated notation of the number p as the sum of the powers of the number 2.

** See the footnote on p. 56.

grouping the summands s_i in the sum (4.19), it is sufficient for the relations

$$S[(x'+x'')+x'''] = S[x'+(x''+x''')]$$

and

$$C[(x'+x'')+x'''] = C[x'+(x''+x''')]$$

to hold true.

The validity of these relations can be established directly, by twice applying formulas (4.5[1]).

We shall now make sure that the functions $S(s)$ and $C(s)$ defined on the set $\{s\}$ possess property 1° on that set. Let s', s'' and $s'+s''$ belong to the indicated set. Let us represent s', s'' and $s'+s''$ as the sum (4.19). Combining the numbers s_n, with the same n entering into s' and s'', until the remaining s_n will have different indices, we shall arrive at the grouping of the summands s_n which gives us the representation (4.19) for the number $s'+s''$. But we have pointed out earlier that the result of computing $S(s)$ and $C(s)$ for the sum of several variables does not depend on the way of grouping the addends of that sum. Consequently, if s', s'' and $s'+s''$ belong to the set $\{s\}$, then the values of $S(s)$ and $C(s)$ computed at those points satisfy the first two relations (4.5[1]). It is easy to verify the validity of the third relation (4.5[1]) for the indicated values of the independent variable. Indeed, it follows from the definition of $S(x)$ and $C(x)$ at points 0 and d that $S^2(0)+C^2(0)=1$ and $S^2(d)+C^2(d)=1$. The recurrence formulas (4.18) yield the validity of the relations $S^2(s_n)+C^2(s_n)=1$ for all s_n, while the formula

$$S^2(x'+x'')+C^2(x'+x'') = (S^2(x')+C^2(x'))(S^2(x'')+C^2(x''))$$

that can be verified directly yields the validity of the relation $S^2(s)+C^2(s)=1$ for all the points of the set $\{s\}$.

Let us now show that for all the points of the set $\{s\}$ different from 0 and d, there hold inequalities

$$0 < S(s) < 1,\ 0 < C(s) < 1.* \tag{4.20}$$

Let us use the method of induction to prove the validity of inequalities (4.20). For that purpose we shall associate every n with a group of members of the set $\{s\}$, placing into this group all the members of $\{s\}$ that can be represented as $pd/2^n$, where $0 < p < 2^n$ and p is an odd number. The elements of this group will be called the elements of order n. Each element of order $n+1$ is contained between two successive elements whose order is not greater than n and which differ from each other by $d/2^n$, i.e. by s_n. The first element of the order $n+1$ is equal to s_{n+1}. All the other elements of order $n+1$ can be obtained by adding to s_{n+1} various s of order n. Let us calculate the values of $S(s_1)$ and $C(s_1)$, with s_1 being the only value of s of the order of unity. From (4.18) we have $S(s_1) = \sqrt{1/2}$ and $C(s_1) = \sqrt{1/2}$. Thus inequalities (4.20) hold true for the elements of the first group. Assume now that inequalities (4.20) hold for all the elements whose order is not higher than n.

Then, according to the first formula (4.5[1]), the values of $S(s)$ are positive at all the points of order $n+1$, and, according to the third formula (4.5[1]), these values are not greater than unity. Putting $x''=d$, $x'=-s$ in the first formula (4.5[1]) and taking into account that the function $C(s)$ is even, we shall find that $C(s)=S(d-s)$, and therefore, for $C(s)$, inequalities (4.20) hold for the values of s of order $n+1$ because of the fact that if s has the order $n+1$, then $d-s$ also has the order $n+1$. It follows by induction that for all the points of the set $\{s\}$ distinct from 0 and d, inequalities (4.20) hold true.

* Recall that at points 0 and d the values of $S(s)$ and $C(s)$ were defined by formulas (4.6[1]).

Let us prove that the functions $S(s)$ and $C(s)$ defined on the set $\{s\}$ are monotonic on that set. We shall show that $S(s)$ is an increasing function and $C(s)$ is a decreasing function. Let $0 \leqslant s' < s'' < d$. Then $\frac{s'+s''}{2}$ and $\frac{s''-s'}{2}$ are contained strictly between zero and d. It follows from formula (4.15) and the inequalities (4.20) that $S(s'') > S(s')$. Consequently, $S(s)$ is an increasing function. From the relation $C(s) = S(d-s)$ it follows that $C(s)$ is a decreasing function on the set $\{s\}$.

Let us prove now that the functions $S(s)$ and $C(s)$ which are defined on the everywhere dense set $\{s\}$ of points of the closed interval $[0, d]$, have a limiting value at each point of the interval $[0, d]$.

First of all we shall consider the sequence $\{s_n\}$ and prove that $\lim\limits_{n\to\infty} S(s_n) = 0$ and $\lim\limits_{n\to\infty} C(s_n) = 1$ (the existence of these limits follows from the monotonicity and boundedness of $S(s)$ and $C(s)$ on the set $\{s\}$). Let us consider the sequence $\left\{\frac{t(s_n)}{s_n}\right\}$, where $t(s_n) = \frac{S(s_n)}{C(s_n)}$. We have from (4.18)

$$2S(s_{n+1})\,C(s_{n+1}) = \sqrt{1-C^2(s_n)} = S(s_n)$$

and

$$C(s_n) = C^2(s_{n+1}) - S^2(s_{n+1}) < C^2(s_{n+1}).$$

Therefore

$$\frac{t(s_n)}{s_n} = \frac{S(s_n)}{s_n C(s_n)} = \frac{2S(s_{n+1})\,C(s_{n+1})}{2s_{n+1}C(s_n)} = \frac{t(s_{n+1})}{s_{n+1}}\,\frac{C^2(s_{n+1})}{C(s_n)} > \frac{t(s_{n+1})}{s_{n+1}}.$$

Thus, $\frac{t(s_n)}{s_n} > \frac{t(s_{n+1})}{s_{n+1}}$ and $\frac{t(s_n)}{s_n} > 0$ for any n, i.e. the sequence $\left\{\frac{t(s_n)}{s_n}\right\}$ is decreasing and bounded. According to Theorem 3.15 it has a limit which we shall designate as L:

$$\lim_{n\to\infty} \frac{t(s_n)}{s_n} = L. \tag{4.21}$$

Since $s_n \to 0$ for $n \to \infty$ we have $\lim\limits_{n\to\infty} t(s_n) = 0$ and, therefore, by virtue of the boundedness of the function $C(s)$ (see (4.20)),

$$\lim_{n\to\infty} S(s_n) = \lim_{n\to\infty} (t(s_n)\,C(s_n)) = 0. \tag{4.22}$$

$C(s)$ being positive, equation (4.22) and relation $S^2(s_n) + C^2(s_n) = 1$ yield

$$\lim_{n\to\infty} C(s_n) = 1. \tag{4.23}$$

Note that (4.21) and (4.23) yield

$$\lim_{n\to\infty} \frac{S(s_n)}{s_n} = L. \tag{4.24}$$

Since $\frac{S(s_n)}{s_n} = \frac{2S(s_{n+1})\,C(s_{n+1})}{2s_{n+1}} < \frac{S(s_{n+1})}{s_{n+1}}$, the sequence $\left\{\frac{S(s_n)}{s_n}\right\}$ increases. Therefore, we have, from (4.21) and (4.24),

$$\frac{S(s_n)}{s_n} < L < \frac{t(s_n)}{s_n}$$

or

$$S(s_n) < L \cdot s_n < t(s_n). \tag{4.25}$$

Assume that $\{s_n^*\}$ is any sequence, converging to zero, of the values s from the set $\{s\}$. It is evidently possible to indicate, for any n, a number k such that $0 < s_n^* < s_k$. Because of the monotonicity of $S(s)$ on the set $\{s\}$, we have $0 < S(s_n^*) < S(s_k)$. Therefore it follows from (4.22) that $\lim_{n\to\infty} S(s_n^*) = 0$.

Let us now prove that the function $S(s)$, defined on the set $\{s\}$*, has a limiting value at any point x of the closed interval $[0, d]$. Suppose $\{s_n'\}$ is a monotone increasing, converging to x, sequence of elements of the set $\{s\}$. Since $\{S(s_n')\}$ is an increasing bounded sequence, there exists a limit $\lim_{n\to\infty} S(s_n')$ which will be denoted by $S(x)$. Assume that $\{s_n''\}$ is any, converging to x, sequence of the elements of the set $\{s\}$ ($s_n'' \neq x$). Then the sequence $\left\{\left|\frac{s_n'' - s_n'}{2}\right|\right\}$ has the limit zero. In accordance with what has been proved, $\lim_{n\to\infty} S\left(\left|\frac{s_n'' - s_n'}{2}\right|\right) = 0$. It follows from (4.15) and the boundedness of the functions $C(s)$ that

$$\lim_{n\to\infty} |S(s_n'') - S(s_n')| = \lim_{n\to\infty} C\left(\frac{s_n'' + s_n'}{2}\right) S\left(\left|\frac{s_n'' - s_n'}{2}\right|\right) = 0.$$

In other words, $\lim_{n\to\infty} S(s_n'') = S(x)$. The sequence $\{s_n''\}$ being arbitrary, this means that there exists a limit of the function $S(s)$, defined on $\{s\}$, at every point x of the closed interval $[0, d]$:

$$\lim_{s\to x} S(s) = S(x).$$

Since the relation $S^2(s) + C^2(s) = 1$ holds true and the function $C(s)$ is nonnegative, on the set $\{s\}$, there exists a limit of the function $C(s)$ at every point of the closed interval $[0, d]$. We shall denote the limit of this function at point x by $C(x)$.

We shall now determine the values of the functions $S(x)$ and $C(x)$ at any point x of the closed interval $[0, d]$ as the limiting values, at point x, of the functions $S(s)$ and $C(s)$ defined on the set $\{s\}$. We shall prove that the functions $S(x)$ and $C(x)$ thus defined possess properties 1° and 2° of the assertion formulated at the beginning of the *proof of the existence of the functions $S(x)$ and $C(x)$*. We shall begin with establishing the fact that the *functions $S(x)$ and $C(x)$* defined on the closed interval $[0, d]$ by the method indicated above are *monotonic and continuous* on that interval. First of all let us prove that if x is any number from the interval $[0, d]$ and s' and s'' are any numbers from the set $\{s\}$ satisfying the inequality $s' < x < s''$, then $S(s') < S(x) < S(s'')$, $C(s') > C(x) > C(s'')$. Let us establish, for instance, the inequality $S(s') < S(x)$ (the inequalities $S(x) < S(s'')$ and $C(s') > C(x) > C(s'')$ can be proved by analogy). Assume that $\{s_n'\}$ is an increasing sequence of numbers from the set $\{s\}$, converging to x, and that all the terms s_n' of the sequence satisfy the inequalities $s' < s_n' < x$. Since the function $S(s)$ increases on the set $\{s\}$, the sequence $\{S(s_n') - S(s')\}$ increases and has positive terms. Consequently, the limit $S(x) - S(s')$** of

* Recall that $\{s\}$ is a set of points dense everywhere in the closed interval $[0, d]$.

** Since $\lim_{n\to\infty} S(s_n') = S(x)$, and $S(s)$ is a fixed number, we have

$$\lim_{n\to\infty} [S(s_n') - S(s')] = S(x) - S(s').$$

this sequence is positive. Thus we have that $S(s') < S(x)$. Let us now prove that the *function $S(x)$ increases on the closed interval* $[0, d]$ (the proof of the *decrease of the function $C(x)$ on that interval* is carried out analogously). Suppose x' and x'' are any two numbers of the interval $[0, d]$ complying with the inequality $x' < x''$. If s' is a number of the set $\{s\}$ contained between x' and x'', $x' < s' < x''$, then, in accordance with what has been proved, $S(x') < S(s')$ and $S(s') < S(x'')$, that is $S(x') < S(x'')$. We have proved the monotonicity of the function $S(x)$ on the closed interval $[0, d]$. *Before starting to prove the continuity of the functions $S(x)$ and $C(x)$, let us ascertain that the limiting values of the functions $S(s)$ and $C(s)$ at points of the set $\{s\}$ coincide with the values of these functions at the corresponding points of the set $\{s\}$.*

Let us consider an arbitrary number s of the set $\{s\}$ and two sequences, $\{s'_n\}$ and $\{s''_n\}$ converging to s, of the elements of the set $\{s\}$, such that $s'_n < s < s''_n$. The inequalities $S(s'_n) < S(s) < S(s''_n)$* hold because of the monotonicity of the function $S(s)$ on the set $\{s\}$. We have proved the formulated assertion because $\lim\limits_{n\to\infty} S(s'_n) = \lim\limits_{n\to\infty} S(s''_n)$ and the indicated limits are equal to the limiting value of the function $S(s)$ at point s. Let us now make sure that the functions $S(x)$ and $C(x)$ are *continuous at each point of the interval* $[0, d]$. For our purpose, it suffices to establish the continuity of these functions at each point x of the indicated interval on the left and on the right, their continuity on the right at point 0 and on the left at point d (see the remark in 4.3.1). Let us prove, for the sake of definiteness, the continuity of the function $S(x)$ at a point x of the interval $[0, d]$ from the left (the continuity on the right and the continuity of $C(x)$ can be proved analogously).

Let $\{s'_k\}$ be a sequence, converging to x on the left, of the numbers of the set $\{s\}$. Since $\lim\limits_{k\to\infty} S(s'_k) = S(x)$, there can be indicated, for any positive ε, a term s'_k of that sequence for which $0 < S(x) - S(s'_k) < \varepsilon$.

Let us discuss now an arbitrary sequence $\{x_n\}$ converging to x on the left. Assume that N is a number beginning with which there hold inequalities $s'_k < x_n < x$. Since the function increases when $n \geqslant N$, the inequalities $S(s'_k) < S(x_n) < S(x)$ hold true. Comparing these inequalities with the inequalities $0 < S(x) - S(s'_k) < \varepsilon$, we see that for $n \geqslant N$ there hold inequalities $0 < S(x) - S(x_n) < \varepsilon$. In other words, the limiting value of the function $S(x)$ at point x on the left is equal to its particular value at that point. Thus we have proved the continuity of $S(x)$ at point x on the left.

We shall now try to define the functions $S(x)$ and $C(x)$ on the interval $[d, 2d]$ with the aid of the relations $S(x + d) = C(x)$ and $C(x + d) = -S(x)$. Applying these formulas once again, we shall extend the functions to the interval $[2d, 4d]$. Repeating the arguments, we shall define the functions for all the positive values of x. To define these functions for the negative values of x, we shall make use of the relations $S(x) = -S(-x)$ and $C(x) = C(-x)$. It is easy to ascertain that as a result we shall obtain the functions which are continuous on the entire number line.

Let us prove that the functions $S(x)$ and $C(x)$ satisfy requirements 1°, 2° and 3° of the assertion formulated at the beginning of the *proof of the existence*. Note that if s', s'', $s' + s''$ and s belong to the set $\{s\}$ of the interval $[0, d]$, then for these values of the argument formulas (4.5[1]) hold true. The method of the extension of the function $S(x)$ and $C(x)$ indicated above implies the validity of these formulas for the values of the argument $d + s'$, s'', where s' and s'' belong to the interval $[0, d]$. Repeating the argument, we shall prove that relations (4.5[1]) are valid for all the values of the argument of the number line of the form $pd/2^n$, where p and n are arbitrary integers. Since these values of the argument

* For definiteness, we prove the assertion for the function $S(x)$.

form a set of points dense everywhere on the number line*, the continuity of the functions $S(x)$ and $C(x)$ implies the validity of relation (4.5[1]) for all values of x.

Since by constructing the functions $S(x)$ and $C(x)$ we have fulfilled requirement 2°, it remains to make sure of the validity of requirement 3°. Let us note the fact that if s', s'' and $s' + s''$ are elements of the set $\{s\}$ of the interval $[0, d]$ and the inequalities $0 < S(s') < Ls'$ and $0 < S(s'') < Ls''$ hold true, then by virtue of the first formula (4.5[1]) and inequalities (4.20) there also hold inequalities $0 < S(s' + s'') < Ls' + Ls'' = L(s' + s'')$. Making use of this remark, formula (4.19) and inequalities (4.20) and (4.25), it is easy to ascertain that the inequalities $0 < S(s) < Ls$ are valid for all s from the set $\{s\}$ on the interval $[0, d]$. Since this set is dense everywhere on $[0, d]$ and $S(x)$ is a continuous function, there hold inequalities $0 < S(x) < Lx$ for all x from $[0, d]$. Thus we have established the validity of requirement 3°.

Now make note of the fact that the number L depends on the choice of d. Namely, if we take the number $d^* = \frac{d}{k}$ instead of d, then $s_n^* = \frac{s_n}{k}$. By construction, $S(s_n^*) = S(s_n)$, and, therefore, $\lim\limits_{n\to\infty} \frac{S(s_n^*)}{s_n^*} = \lim\limits_{n\to\infty} k \frac{S(s_n)}{s_n} = kL$ (see (4.24)). Choosing $k = \frac{1}{L}$, we shall define on the interval $[0, d^*]$ the functions $S(x)$ and $C(x)$ such that the inequalities $0 < S(x) < x$ will hold.

Geometrical considerations show that if $d = \frac{\pi}{2}$, then $2S(s_n)$ is the length of the side of a regular 2^n-gon inscribed into a circle of radius 1, $2s_n$ is the length of the arc of a circle subtended by a chord $2S(s_n)$ in length, and $2t(s_n)$ is the length of the side of a regular 2^n-gon circumscribed about that circle. In that case, inequalities (4.25) have the form $S(s_n) < s_n < t(s_n)$. Therefore, in the case discussed, $L = 1$. The proof of the assertion is complete.

---

* See footnote on p. 140.

CHAPTER 5

FUNDAMENTALS OF DIFFERENTIAL CALCULUS

In this chapter we introduce the concepts of a derivative and a differential, establish the rules of differentiation, calculate the derivatives of all the basic elementary functions described in Chap. 1. Then follow the derivatives and differentials of higher orders.

5.1. A DERIVATIVE. ITS PHYSICAL AND GEOMETRICAL INTERPRETATIONS

5.1.1. Increment of an argument and of a function. Difference form of the condition of continuity. Suppose the function $y = f(x)$ is defined on a certain open interval* (a, b). We shall fix any value x from the indicated interval and assign to the argument at point x an arbitrary increment Δx such that the value $x + \Delta x$ also belongs to the interval (a, b). *The increment of the function* $y = f(x)$ *at point* x *corresponding to the increment of the argument* Δx *is said to be the number*

$$\Delta y = f(x + \Delta x) - f(x). \tag{5.1}$$

Thus, for the function $y = \sin x$, the increment at point x, corresponding to the increment of the argument Δx, is equal to

$$\Delta y = \sin(x + \Delta x) - \sin x = 2\cos\left(x + \frac{\Delta x}{2}\right)\sin\frac{\Delta x}{2}. \tag{5.2}$$

It is proper to make here the following assertion: *for the function* $y = f(x)$ *to be continuous at point* x *it is necessary and sufficient for the increment* Δy *of that function at point* x, *corresponding to the increment of the argument* Δx, *to be infinitesimal at* $\Delta x \to 0$.

Indeed, by definition, the function $y = f(x)$ is continuous at point x if there exists a limiting value

$$\lim_{\Delta x \to 0} f(x + \Delta x) = f(x). \tag{5.3}$$

* Instead of the open interval (a, b) we can consider a closed interval $[a, b]$, a half-line or an infinite interval, or, generally, *any* set $\{x\}$ *dense in itself*. The definition of the set $\{x\}$ dense in itself was given in 2.3.

In accordance with 4.2.3, the existence of the limiting value (5.3) is equivalent to the function $[f(x+\Delta x)-f(x)]$ of the argument Δx being infinitesimal at $\Delta x \to 0$.

The assertion we have proved makes it possible to express the condition of continuity of the function $y=f(x)$ at point x in a new form, namely, *the function $y=f(x)$ is continuous at point x* if the increment Δy *of that function at point x, corresponding to the increment of the argument Δx, is infinitesimal at $\Delta x \to 0$*, i.e. if

$$\lim_{\Delta x \to 0} \Delta y = \lim_{\Delta x \to 0} [f(x+\Delta x)-f(x)] = 0. \tag{5.4}$$

Condition (5.4) will be called the *difference form of the condition of continuity of the function $y=f(x)$ at point x*. This condition will be repeatedly used in what follows.

We shall use condition (5.4) to make sure once again that the function $y=\sin x$ is continuous at any point x of an infinite interval. Indeed, it follows directly from formula (5.2), condition $\left|\cos\left(x+\frac{\Delta x}{2}\right)\right| \leqslant 1$ and equality $\lim\limits_{\Delta x \to 0} \sin\frac{\Delta x}{2}=0$ that $\lim\limits_{\Delta x \to 0} \Delta y = 0$.

5.1.2. Definition of a derivative. We shall retain for the function $y=f(x)$ the assumptions and designations formulated at the beginning of 5.1.1.

Postulating that $\Delta x \neq 0$, we shall consider at a given *fixed* point x the ratio of the increment Δy of the function at that point to the corresponding increment of the argument Δx

$$\frac{\Delta y}{\Delta x} = \frac{f(x+\Delta x)-f(x)}{\Delta x}. \tag{5.5}$$

Expression (5.5) will be called the *difference quotient* (at the given point x). Since the value of x is assumed to be *fixed*, difference quotient (5.5) is the *function of the argument* Δx. This function is defined for all the values of the argument Δx belonging to a certain infinitesimal neighbourhood of point $\Delta x = 0$, with the exception of the point $\Delta x = 0$ itself. Thus we can rightfully consider the question of the existence of a limit of the indicated function at $\Delta x \to 0$.

Definition. *We shall use the term the* **derivative** *of the function $y=f(x)$ at a given fixed point x for the limit, at $\Delta x \to 0$, of the difference quotient* (5.5) *(under the condition that the limit does exist).*

The derivative of the function $y=f(x)$ at point x will be designated as $y'(x)$ or $f'(x)$. Thus, according to the definition,

$$f'(x) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x+\Delta x)-f(x)}{\Delta x}. \tag{5.6}$$

It might be well to point out that if the function $y=f(x)$ is defined and has a derivative for all x from the interval (a, b), then the deriv-

ative will be some function of the variable x, also defined on the interval (a, b).

5.1.3. Physical interpretation of a derivative. In Chap. 1 the concept of a derivative was discussed in terms of physics. Here again we shall deal with the physical applications of the concept of a derivative.

Let us assume, first of all, that the function $y = f(x)$ describes the *law of motion of a particle in a straight line**. Then, as it is known, the difference quotient (5.5) is the expression of the mean velocity of the particle in the time interval from x to $x + \Delta x$. In that case, the derivative $f'(x)$, i.e. the limit of the difference quotient (5.5), at $\Delta x \to 0$, determines the *instantaneous velocity of the particle at the time moment* x. Thus we see that the derivative of the function describing the law of motion determines the instantaneous velocity of the particle.

To obviate the wrong inference that the concept of a derivative is widely used only in mechanics, we shall give examples of the application of this concept in other branches of physics.

Let us assume that the function $y = f(x)$ determines the electric current y flowing through the cross section of a conductor during the time x. (The time moment $x = 0$ is taken to be the reference point.) In that case the derivative $f'(x)$ will determine the *intensity of current* passing through the cross section of the conductor at the time moment x.

Let us proceed considering the process of heating of some body. Assume that the function $y = f(x)$ determines the amount of heat** y which must be imparted to the body to raise its temperature from $0°$ to $x°$. Then, as is known from elementary physics, the difference quotient (5.5) determines the *average heat capacity* of the body being heated from $x°$ to $(x + \Delta x)°$. In this case the derivative $f'(x)$, i.e. the limiting value of the difference quotient (5.5), at $\Delta x \to 0$, determines the *heat capacity of the body at the given temperature* x. It might be well to emphasize that this heat capacity, generally speaking, changes with a change in temperature x.

We have discussed the applications of the concept of a derivative in three different branches of physics. In the course of general physics the reader will encounter various other examples of application of the concept of a derivative.

5.1.4. Geometrical interpretation of a derivative. We have considered in 1.2 the problem of finding a tangent to a curve which is the graph of the function $y = f(x)$ (on a certain interval (a, b)). We have given there the definition of a tangent to the indicated

---

* That is, the dependence of the path y traversed by the particle from the origin on time x.

** Expressed in calories, for instance.

curve at a point $M(x, f(x))$ of that curve. (Here x is the value of the independent variable from the interval (a, b). See Fig. 5.1.) If we designate by Δx an arbitrary increment of the variable and by P a point of the curve with the coordinates $(x + \Delta x, f(x + \Delta x))$, then we shall define the tangent passing through point M of that curve as the limiting position of the secant MP at Δx tending to

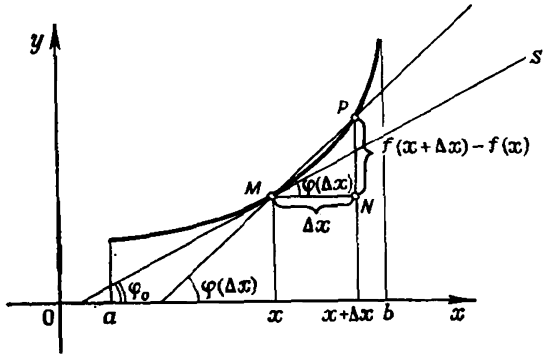


Fig. 5.1

zero. It is clear from Fig. 5.1 that the slope of the secant MP (i.e. the slope of the secant to the x-axis) is equal to the difference quotient (5.5). From this fact and from the fact that in the limit, when Δx tends to zero, the slope of the secant must pass into the slope of the tangent, we have made a conclusion, in 1.2 based on what we saw, in the figure, that the *derivative $f'(x)$ is equal to the slope of the tangent at point M to the graph of the function $y = f(x)$*.

Here we shall elucidate the indicated visual considerations. Assuming that the function $y = f(x)$ has a derivative at a given point x, we shall prove that (1) the graph of the function $y = f(x)$ has a tangent at a given point $M(x, f(x))$, and (2) the slope of the indicated tangent is equal to $f'(x)$.

We shall prove assertions (1) and (2) simultaneously. Let us designate the slope of the secant MP to the x-axis by the symbol $\varphi(\Delta x)$. Since the slope of the secant MP (i.e. $\tan \varphi(\Delta x)$) is equal to the ratio $\Delta y/\Delta x$, we have

$$\varphi(\Delta x) = \arctan \frac{\Delta y}{\Delta x} \tag{5.7}$$

for any sufficiently small Δx different from zero. The existence of a derivative $f'(x)$, that is the existence of a limiting value $\lim_{\Delta x \to 0} (\Delta y/\Delta x) = f'(x)$ and the continuity of the function $u = \arctan x$ for all the values of the argument imply the existence of a limiting value of the function (5.7) at point $\Delta x = 0$ and of the

equality

$$\lim_{\Delta x \to 0} \varphi(\Delta x) = \lim_{\Delta x \to 0} \operatorname{arc\,tan} \frac{\Delta y}{\Delta x} = \operatorname{arc\,tan} f'(x). \tag{5.8}$$

Equality (5.8) proves the existence of a limiting value (for $\Delta x \to 0$) of the slope of the secant MP, i.e. proves the existence of a tangent at point M. Besides, equality (5.8) implies that if the slope of the tangent is denoted by φ_0, then $\varphi_0 = \operatorname{arc\,tan} f'(x)$, that is $\tan \varphi_0 = f'(x)$.

5.1.5. Right-hand and left-hand derivatives. The concepts of the *right-hand (forward)* and *left-hand (backward) derivatives* of the function $y = f(x)$ at a given fixed point x are introduced by a complete analogy with the right-hand and left-hand limits of a function.

Definition. *We use the term the* **right-hand (left-hand) derivative** *of the function $y = f(x)$ at a given fixed point x for the right-hand (left-hand) limiting value of the difference quotient* (5.5) *at point $\Delta x = 0$ (under the condition that this limiting value does exist).*

It is usual to designate the right-hand derivative of the function $y = f(x)$ at point x as $f'(x + 0)$, and the left-hand derivative at point x as $f'(x - 0)$.

*If the function $y = f(x)$ has a derivative at point x, it has both the right-hand and the left-hand derivative at that point and those derivatives coincide. If the function $y = f(x)$ has both the right-hand and the left-hand derivative at point x and the derivatives coincide, then the function $y = f(x)$ has a derivative at point x**. However, there are functions which have the right-hand and the left-hand derivative at point x but have no derivative at that point. We can cite as an example the function

$$f(x) = |x| = \begin{cases} +x, & \text{if } x \geqslant 0, \\ -x, & \text{if } x < 0. \end{cases}$$

This function has, at point $x = 0$, a right-hand derivative equal to $\lim\limits_{\Delta x \to 0+0} \frac{\Delta x}{\Delta x} = 1$ and a left-hand derivative equal to $\lim\limits_{\Delta x \to 0-0} \frac{-\Delta x}{\Delta x} = -1$, but has no derivative at point $x = 0$.

5.1.6. The concept of a derivative of a vector function. In mathematical analysis and its applications the reader will often come across the concepts of a vector function and its derivative.

If every value of the variable t from some set $\{t\}$ is associated, in accordance with a certain law, with a definite vector **a**, *then it is said that a vector function* $\mathbf{a} = \mathbf{a}(t)$ *is defined on the set $\{t\}$.*

Since every vector **a** is unambiguously determined, in the Cartesian rectangular system, by the three coordinates x, y and z, the definition of the vector

* This assertion follows from the corresponding assertion for the right-hand and left-hand limits of a function (see the remark in 4.2.1).

function $\mathbf{a} = \mathbf{a}(t)$ is equivalent to the definition of three scalar functions $x = x(t)$, $y = y(t)$ and $z = z(t)$.

We can visualize the concept of a vector function more vividly if we consider the so-called *hodograph* of that function.

A hodograph is a locus of the end-points of all vectors $\mathbf{a}(t)$ issuing from the origin O. The curve L in Fig. 5.2 is a hodograph of the vector function $\mathbf{a} = \mathbf{a}(t)$.

The concept of a hodograph of a vector function is a generalization of the concept of a graph of a scalar function.

Let us introduce the concept of a derivative of the vector function $\mathbf{a}(t)$ at a given fixed point t. For that purpose, we shall give the argument t an arbitrary increment $\Delta t \neq 0$ and consider the vector $\Delta\mathbf{a} = \mathbf{a}(t + \Delta t) - \mathbf{a}(t)$ (in Fig. 5.2 this vector coincides with the vector $\overrightarrow{MP}$). Having multiplied the indicated vector by the number $1/\Delta t$, we obtain a new vector

$$\frac{\Delta a}{\Delta t} = \frac{1}{\Delta t}[\mathbf{a}(t+\Delta t) - \mathbf{a}(t)], \qquad (5.5^*)$$

collinear with the original vector. Vector (5.5*) is an analogue of the difference quotient (5.5). Note that vector (5.5*) represents the average rate of change of a vector function on the closed interval $[t, t + \Delta t]$.

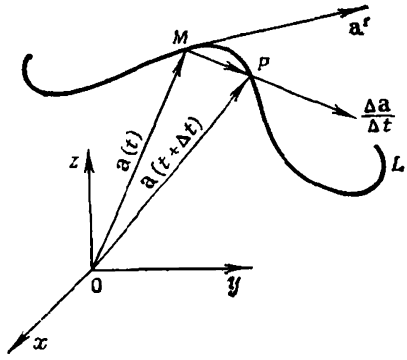


Fig. 5.2

The derivative of the vector function $\mathbf{a} = \mathbf{a}(t)$ *at a given fixed point* t *is said to be the limit of the difference quotient* (5.5*) *at* $\Delta t \to 0$.

The derivative of the vector function $\mathbf{a}(t)$ is symbolized as $\mathbf{a}'(t)$ or $d\mathbf{a}/dt$.

It is evident from geometrical considerations that the derivative of the vector function $\mathbf{a} = \mathbf{a}(t)$ is a vector which is tangent to the hodograph of that function.

Since the coordinates of the difference quotient (5.5*) are equal, respectively, to

$$\frac{x(t+\Delta t) - x(t)}{\Delta t}, \quad \frac{y(t+\Delta t) - y(t)}{\Delta t}, \quad \frac{z(t+\Delta t) - z(t)}{\Delta t},$$

it is clear that the coordinates of the derivative $\mathbf{a}'(t)$ are equal to the derivatives of the functions $x'(t)$, $y'(t)$ and $z'(t)$. Thus, the computation of the derivative of a vector function reduces to the computation of the derivatives of its coordinates.

Remark 1. Since the vector function $\mathbf{a} = \mathbf{a}(t)$ specifies the law of motion of a particle along a curve L representing the hodograph of that function, the derivative $\mathbf{a}'(t)$ is equal to the rate of motion along the indicated curve.

Remark 2. It is known from elementary geometry that vector products may be of different types: scalar product, vector product and mixed product. If we express all these products in coordinates, we shall be able to indicate the rules governing the computations of the derivatives of the corresponding products of vector functions. By way of example we will give the rule governing the computation of the derivative of the scalar product of two vector functions $\mathbf{a}(t) = \{a_1(t), a_2(t), a_3(t)\}$ and $\mathbf{b}(t) = \{b_1(t), b_2(t), b_3(t)\}$: $\{\mathbf{a}(t)\,\mathbf{b}(t)\}' = \mathbf{a}'(t)\,\mathbf{b}(t) + \mathbf{a}(t)\,\mathbf{b}'(t) = \{a_1'(t)\,b_1(t) + a_2'(t)\,b_2(t) + a_3'(t)\,b_3(t)\} + \{a_1(t)\,b_1'(t) + a_2(t)\,b_2'(t) + a_3(t)\,b_3'(t)\}$. A similar rule is also valid for the vector product of two vector functions

$$[\mathbf{a}(t)\,\mathbf{b}(t)]' = [\mathbf{a}'(t)\,\mathbf{b}(t)] + [\mathbf{a}(t)\,\mathbf{b}'(t)].$$

5.2. THE CONCEPT OF A DIFFERENTIABLE FUNCTION

5.2.1. Differentiation of a function at a given point. We shall assume that, as in 5.5.1 and 5.1.2, the function $y = f(x)$ is defined in a certain interval (a, b), the symbol x is used to denote a certain fixed value of the argument from the interval indicated, and the symbol Δx is employed to denote any increment of the argument such that the value of the argument $x + \Delta x$ also belongs to (a, b).

Definition. *The function $y = f(x)$ is said to be* **differentiable** *at a given point x if the increment Δy of that function at point x, corresponding to the increment of the argument Δx, can be represented as*

$$\Delta y = A\Delta x + \alpha\Delta x, \tag{5.9}$$

where A is some number independent of Δx, and α is a function of the variable Δx, infinitesimal in the case of Δx tending to zero.

Make a note of the fact that the function $\alpha(\Delta x)$ can assume *any value* at point $\Delta x = 0$ (in this case representation (5.9) remains valid at that point). For the sake of definiteness we can put $\alpha(0) = 0$*.

Since the product of two infinitesimals $\alpha\Delta x$ is an infinitesimal of a higher order than Δx (see 4.2.3), i.e. $\alpha\Delta x = o(\Delta x)$, formula (5.9) can be rewritten in the form

$$\Delta y = A\Delta x + o(\Delta x).$$

Theorem 5.1. *For the function $y = f(x)$ to be differentiable at a given point x, it is necessary and sufficient for it to have a finite derivative at that point.*

Proof. (1) **Necessity.** Let the function $y = f(x)$ be differentiable at a given point x, that is its increment Δy at that point be representable in the form (5.9). Assume that $\Delta x \neq 0$ and divide equality (5.9) by Δx. The result will be

$$\frac{\Delta y}{\Delta x} = A + \alpha. \tag{5.10}$$

Equality (5.10) suggests the existence of a derivative, i.e. of the limiting value $\lim\limits_{\Delta x \to 0} (\Delta y/\Delta x) = A$.

(2) **Sufficiency.** Let the function $y = f(x)$ have a finite derivative at a given point x, i.e. assume that there exists a limit

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = f'(x). \tag{5.11}$$

By virtue of the definition of the limiting value, the function $\alpha = \frac{\Delta y}{\Delta x} - f'(x)$ of the argument Δx is an infinitesimal in the case of

* In this case the particular value of the function $\alpha(\Delta x)$ at point $\Delta x = 0$ coincides with its limiting value at that point.

Δx tending to zero, i.e.

$$\Delta y = f'(x)\,\Delta x + \alpha\Delta x, \tag{5.12}$$

where $\lim\limits_{\Delta x \to 0} \alpha = 0$. Representation (5.12) coincides with representation (5.9) if the number $f'(x)$ independent of Δx is denoted by A. Thus we have proved that the function $y = f(x)$ is differentiable at point x. That was the proof of Theorem 5.1.

The theorem we have proved makes it possible to *identify*, in what follows, *the notion of the differentiability of a function at a given point with the notion of the existence of a derivative of the function at that point.*

From now on the operation of finding a derivative will be called *differentiation.*

5.2.2. The connection between the concepts of differentiability and continuity of a function. We can make the following elementary assertion.

Theorem 5.2. *If the function $y = f(x)$ is differentiable at a given point x, it is automatically continuous at that point.*

Proof. Since the function $y = f(x)$ is differentiable at point x, its increment Δy at that point can be represented in the form (5.9). But it follows from formula (5.9) that $\lim\limits_{\Delta x \to 0} \Delta y = 0$, i.e. that the function $y = f(x)$ is continuous at point x because of the difference form of the condition of continuity (see 1.1). Thus, we have proved the theorem.

The question naturally arises as to the validity of the assertion converse of Theorem 5.2, i.e. whether the continuity of a function at a given point means its differentiability at that point. The answer is negative, for there axist functions continuous at a certain point without being differentiable at that point. The function $y = |x|$ can serve as an example. It is evident that this function is continuous at point $x = 0$ but, as shown at the end of 1.5, it is not differentiable at that point. Make a note of the fact that there exist functions continuous on a certain closed interval but having no derivatives at any point of that interval*.

5.2.3. The concept of a differential of a function. Assume that the function $y = f(x)$ is differentiable at a point x, that is the increment Δy of that function at point x can be written in the form (5.9). Analysing formula (5.9) we infer that the increment Δy of the differentiated function is the *sum of two addends*: the first of the addends $A\Delta x$, at $A \neq 0$, is the function of the increment Δx of the argument,

* The first example of such a function was published by Weierstrass. Somewhat earlier a similar example was obtained by the Czech mathematician Bolzano, but the example was not published. Such a function will be considered in the supplement to Chap. 11.

linear and homogeneous with respect to* Δx; when $\Delta x \to 0$, this addend is an *infinitesimal of the same order as* Δx; the *second* addend $\alpha \Delta x$ is an *infinitesimal of the higher order than* Δx for $\Delta x \to 0$, since the ratio $(\alpha \Delta x / \Delta x) = \alpha$ tends to zero when $\Delta x \to 0$. Thus, for $A \neq 0$, the first addend $A \Delta x$ is the *principal part* of the increment of the differentiable function. This principal part of the increment is called the differential of the function at point x corresponding to the increment of the argument Δx.

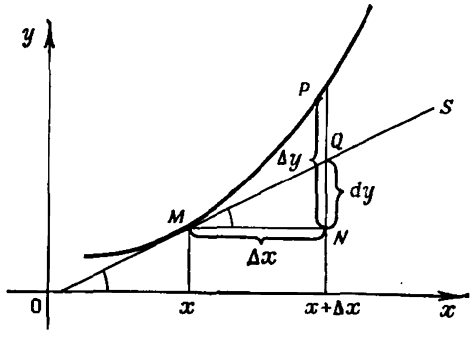


Fig. 5.3

Thus, in the case $A \neq 0$, *the differential of the function* $y=f(x)$ *at a given point* x, *corresponding to the increment of the argument* Δx, *is the principal, linear with respect to* Δx, *part of the increment of that function at point* x. It is customary to denote the differential of the function $y=f(x)$ as dy. If representation (5.9) is valid for the increment of the function Δy, then, by definition, the differential of this function is equal to

$$dy = A \cdot \Delta x. \tag{5.13}$$

When A is equal to zero, the summand $A \cdot \Delta x$ is no longer the principal part of the increment Δy of the differentiated function (for this summand is equal to zero while the summand $\alpha \cdot \Delta x$ is, in general, different from zero). However, it is usually agreed, in the case $A = 0$, to specify the differential of the function by formula (5.13), i.e. to consider it to be equal to zero in this case.

If we take Theorem 5.1 into account, i.e. assume that $A = f'(x)$, then formula (5.13) can be rewritten as

$$dy = f'(x)\, \Delta x. \tag{5.14}$$

Formula (5.14) expresses the differential of the function at point x, corresponding to the increment of the argument Δx. It should be pointed out that the differential of the function dy at the given point x is not equal, in general, to the increment of the function Δy at that point. This becomes especially clear when we consider the graph of the function $y = f(x)$ (Fig. 5.3). Let a point M on the curve $y = f(x)$ correspond to the value of the argument x, a point P on the same curve correspond to the value of the argument $x + \Delta x$, and MS be a tangent to the curve $y = f(x)$ at point M. Further,

* Recall that the *linear function* of the argument x is a function of the form $y = Ax + B$, where A and B are some constants. In the case $B = 0$ the linear function is said to be *homogeneous*.

let $MN \parallel Ox$, $PN \parallel Oy$, Q being the point of intersection of the tangent MS and the line PN. Then, the increment of the function Δy is equal to the value of the segment NP. At the same time, it is clear from the right-angled triangle MQN and formula (5.14) that the differential of the function dy is equal to the value of the segment NQ, for the value of the segment MN is equal to Δx, and the tangent of the angle QMN is equal to $f'(x)$. It is evident that, in a general case, the values of the segments NP and NQ are different.

To conclude this subsection, we shall determine the expression for the differential of the function $y = f(x)$, whose argument x is an *independent variable**.

We shall now introduce the concept of a *differential dx of the independent variable x*. Any number, independent of x, can be taken to be the differential dx of the independent variable x. Let us agree, from now on, to take this number equal to the increment Δx of the independent variable**. This agreement enables us to rewrite formula (5.14) in the form

$$dy = f'(x)\,dx. \tag{5.15}$$

We want to emphasize that formula (5.15) has as yet been established only for the case when the argument x is an *independent variable*. However, later on, in 5.9, we shall prove that formula (5.15) remains valid for the case when the argument x is not an independent variable but is itself a differentiable function of some new variable.

For the time being, we can make the following conclusion from formula (5.15); in the case when the argument x of the function $y = f(x)$ is an independent variable, the derivative $f'(x)$ of that function is equal to the ratio of the differential of the function dy to the differential of the argument dx, i.e.

$$f'(x) = \frac{dy}{dx}.$$

It will be shown in 5.9 that this ratio is also valid in the case when the argument x is itself a differentiable function of some new variable.

5.3. RULES FOR DIFFERENTIATING A SUM, A DIFFERENCE, A PRODUCT AND A QUOTIENT

Theorem 5.3. *If each of the functions $u(x)$ and $v(x)$ is differentiable at a given point x, then, the sum, the difference, the product and the quotient of these functions (in the case of a quotient the condition*

* We should like to point out that the argument x of the function $y = f(x)$ can, generally speaking, be itself a function of some variable.

** This agreement can be justified if we consider the independent variable x as a function of the form $y = x$, for which $dy = dx = \Delta x$.

$v(x) \neq 0$ *must be fulfilled) are also differentiable at that point, there holding the equalities*

$$[u(x) \pm v(x)]' = u'(x) \pm v'(x),$$
$$[u(x)\,v(x)]' = u'(x)\,v(x) + u(x)\,v'(x), \tag{5.16}$$
$$\left[\frac{u(x)}{v(x)}\right]' = \frac{u'(x)\,v(x) - u(x)\,v'(x)}{v^2(x)}.$$

Proof. We shall consider separately the cases of a sum (difference), a product and a quotient.

1°. Assume $y(x) = u(x) \pm v(x)$. Let us symbolize as Δu, Δv and Δy the increments of the functions $u(x)$, $v(x)$ and $y(x)$ at a given point x corresponding to the increment of the argument Δx. Then evidently

$$\Delta y = y(x + \Delta x) - y(x) = [u(x + \Delta x) \pm v(x + \Delta x)] -$$
$$- [u(x) \pm v(x)] = [u(x + \Delta x) - u(x)] \pm$$
$$\pm [v(x + \Delta x) - v(x)] = \Delta u \pm \Delta v.$$

Thus, for $\Delta x \neq 0$

$$\frac{\Delta y}{\Delta x} = \frac{\Delta u}{\Delta x} \pm \frac{\Delta v}{\Delta x}. \tag{5.17}$$

Now let Δx tend to zero. Then, because of the existence of the derivatives of the functions $u(x)$ and $v(x)$ at point x, there is a limiting value of the right-hand part of (5.17) equal to $u'(x) \pm v'(x)$. Consequently, there must exist a limiting value (for $\Delta x \to 0$) of the left-hand side of (5.17) as well. According to the definition of the derivative, the indicated limiting value is equal to $y'(x)$, and we arrive at the required equality

$$y'(x) = u'(x) \pm v'(x).$$

2°. Next, let us have $y(x) = u(x)\,v(x)$. Retaining for Δu, Δv and Δy the same meaning as before, we shall have

$$\Delta y = y(x + \Delta x) - y(x) = u(x + \Delta x)\,v(x + \Delta x) -$$
$$- u(x)\,v(x) = [u(x + \Delta x)\,v(x + \Delta x) - u(x + \Delta x)\,v(x)] +$$
$$+ [u(x + \Delta x)\,v(x) - u(x)\,v(x)]$$

(we have added and subtracted the term $u(x + \Delta x)\,v(x)$). We can further write

$$\Delta y = u(x + \Delta x)\,[v(x + \Delta x) - v(x)] + v(x)\,[u(x + \Delta x) - u(x)] =$$
$$= u(x + \Delta x)\,\Delta v + v(x)\,\Delta u.$$

Thus, for $\Delta x \neq 0$, we have

$$\frac{\Delta y}{\Delta x} = u(x + \Delta x)\,\frac{\Delta v}{\Delta x} + v(x)\,\frac{\Delta u}{\Delta x}. \tag{5.18}$$

Now let Δx tend to zero. Then, by virtue of the differentiability of the functions $u(x)$ and $v(x)$ at point x, there exist limiting value of the ratios $\Delta u/\Delta x$ and $\Delta v/\Delta x$ equal, respectively, to $u'(x)$ and $v'(x)$. Further, in accordance with Theorem 5.2, the differentiability of $u(x)$ at point x implies the continuity of $u(x)$ at that point. Consequently, there exists a limiting value $\lim\limits_{\Delta x \to 0} u(x + \Delta x)$ equal to $u(x)$. Thus, when Δx tends to zero, the right-hand side of (5.18) has a limit equal to $u(x) v'(x) + v(x) u'(x)$. This means that the left-hand side of (5.18) also has a limit (for $\Delta x \to 0$). By the definition of a derivative, the indicated limit is equal to $y'(x)$, and we arrive at the required formula

$$y'(x) = u'(x) v(x) + u(x) v'(x).$$

3°. Let us, finally, assume that $y(x) = u(x)/v(x)$. Then*

$$\Delta y = y(x+\Delta x) - y(x) = \frac{u(x+\Delta x)}{v(x+\Delta x)} - \frac{u(x)}{v(x)} =$$

$$= \frac{u(x+\Delta x)\, v(x) - v(x+\Delta x)\, u(x)}{v(x)\, v(x+\Delta x)}.$$

Adding and subtracting the summand $u(x) v(x)$ in the numerator, we obtain

$$\Delta y = \frac{[u(x+\Delta x) v(x) - u(x) v(x)] - [v(x+\Delta x) u(x) - u(x) v(x)]}{v(x) v(x+\Delta x)} =$$

$$= \frac{v(x)[u(x+\Delta x) - u(x)] - u(x)[v(x+\Delta x) - v(x)]}{v(x) v(x+\Delta x)} = \frac{v(x)\Delta u - u(x)\Delta v}{v(x) v(x+\Delta x)}.$$

Thus we have, for $\Delta x \neq 0$,

$$\frac{\Delta y}{\Delta x} = \frac{v(x)\dfrac{\Delta u}{\Delta x} - u(x)\dfrac{\Delta v}{\Delta x}}{v(x)\, v(x+\Delta x)}. \tag{5.19}$$

Assume now that Δx tends to zero. By virtue of the differentiability (and the consequent continuity) of the functions $u(x)$ and $v(x)$ at point x, there exist limiting values

$$\lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} = u'(x), \quad \lim_{\Delta x \to 0} \frac{\Delta v}{\Delta x} = v'(x), \quad \lim_{\Delta x \to 0} v(x+\Delta x) = v(x).$$

Thus, because of $v(x) \neq 0$, the right-hand side of (5.19) has a limit, for $\Delta x \to 0$, equal to

$$\frac{v(x) u'(x) - v'(x) u(x)}{v^2(x)}.$$

* Since in what follows the denominator will contain the value $v(x + \Delta x)$, we must prove that this value is different from zero *for all sufficiently small* Δx. Indeed, were it not so, there will be an *infinitesimal* sequence of values Δx_n such that $v(x + \Delta x_n) = 0$. But inasmuch as the function $v(x)$ is continuous for the value of the argument x, the condition $v(x + \Delta x_n) = 0$ would imply $v(x) = 0$, and that contradicts the hypothesis.

And this means that the left-hand side of (5.19) also has a limit (for $\Delta x \to 0$). In accordance with the definition of a derivative, the indicated limiting value is equal to $y'(x)$, and we obtain the required formula

$$y'(x) = \frac{u'(x)\,v(x) - v'(x)\,u(x)}{v^2(x)}.$$

The proof of Theorem 5.3 is complete.

5.4. CALCULATING THE DERIVATIVES OF A POWER FUNCTION, TRIGONOMETRIC FUNCTIONS AND A LOGARITHMIC FUNCTION

In this section we shall deal with the derivatives of basic elementary functions.

5.4.1. The derivative of a power function with an integral-valued exponent. We shall begin with the calculation of the derivative of the power function $y = x^n$, whose exponent n is a positive *integer**. The case of a power function whose exponent is *any real number*, not necessarily integral, will be treated in 5.8.

Making use of Newton's binomial formula, we can write

$$\Delta y = (x+\Delta x)^n - x^n = \left[x^n + nx^{n-1}\Delta x + \frac{n(n-1)}{2}x^{n-2}(\Delta x)^2 + \ldots\right.$$

$$\left. \ldots + (\Delta x)^n\right] - x^n = nx^{n-1}\Delta x + \frac{n(n-1)}{2}x^{n-2}(\Delta x)^2 + \ldots + (\Delta x)^n.$$

Thus, for $\Delta x \neq 0$,

$$\frac{\Delta y}{\Delta x} = nx^{n-1} + \frac{n(n-1)}{2}x^{n-2}\Delta x + \ldots + (\Delta x)^{n-1}. \tag{5.20}$$

Since all the summands on the right-hand side of equation (5.20), beginning with the second onward, contain Δx in positive powers as a factor, there exists a limiting value of the indicated summands, for $\Delta x \to 0$ equal to zero. The first summand on the right-hand side of (5.20) does not depend on Δx. Consequently, there exists a limiting value (for $\Delta x \to 0$) of the right-hand side of (5.20) equal to nx^{n-1}. According to the definition of a derivative, the indicated limiting value is equal to the derivative of the function $y = x^n$, i.e.

$$(x^n)' = nx^{n-1}.$$

This argument is valid for any point x of the infinite interval.

5.4.2. The derivative of the function $y = \sin x$. Using the formula which makes possible the reduction of the difference between the sines to a form convenient for taking logarithms, we can write

$$\Delta y = \sin(x+\Delta x) - \sin x = 2\cos\left(x + \frac{\Delta x}{2}\right)\sin\frac{\Delta x}{2}.$$

* This derivative was treated in Chap. 1 by means of the intuitive approach to the limit.

Thus, for $\Delta x \neq 0$

$$\frac{\Delta y}{\Delta x} = \cos\left(x + \frac{\Delta x}{2}\right) \frac{\sin\left(\frac{\Delta x}{2}\right)}{\left(\frac{\Delta x}{2}\right)}. \tag{5.21}$$

Since the function $y = \cos x$ is continuous at any point x of the infinite interval*, there exists a limiting value

$$\lim_{\Delta x \to 0} \cos\left(x + \frac{\Delta x}{2}\right) = \cos x. \tag{5.22}$$

Further, by virtue of the main result obtained in Sec. 4.6.2, there exists a limiting value

$$\lim_{\Delta x \to 0} \frac{\sin\left(\frac{\Delta x}{2}\right)}{\left(\frac{\Delta x}{2}\right)} = 1. \tag{5.23}$$

Thus, there exists a limiting value (for $\Delta x \to 0$) of the right-hand side of (5.21) equal to the product of the limiting values of (5.22) and (5.23), that is equal to $\cos x$. In accordance with the definition of a derivative, the indicated limiting value is equal to the derivative of the function $y = \sin x$, i.e.

$$(\sin x)' = \cos x.$$

This argument is valid for any point x of the infinite interval.

5.4.3. The derivative of the function $y = \cos x$. Making use of the formula which reduces the difference between the cosines to the form convenient for taking logarithms, we can write

$$\Delta y = \cos(x + \Delta x) - \cos x = -2 \sin\left(x + \frac{\Delta x}{2}\right) \sin\frac{\Delta x}{2}.$$

Thus, for $\Delta x \neq 0$

$$\frac{\Delta y}{\Delta x} = -\sin\left(x + \frac{\Delta x}{2}\right) \frac{\sin\left(\frac{\Delta x}{2}\right)}{\left(\frac{\Delta x}{2}\right)}. \tag{5.24}$$

Since the function $y = \sin x$ is continuous at any point x of the infinite interval, there is a limiting value

$$\lim_{\Delta x \to 0} \sin\left(x + \frac{\Delta x}{2}\right) = \sin x. \tag{5.25}$$

* This was proved in 4.5.6. By the way, the continuity of the function $y = \cos x$ can be easily verified by means of the *difference form* of the condition of continuity.

The existence of the limiting values (5.23) and (5.25) presupposes the existence of the limit (for $\Delta x \to 0$) of the right-hand side of (5.24) equal to $(-\sin x)$. By the definition of a derivative the last limiting value is equal to the derivative of the function $y = \cos x$, i.e.

$$(\cos x)' = -\sin x.$$

The argument is valid for any point x of the infinite interval.

5.4.4. The derivatives of the functions $y = \tan x$ and $y = \cot x$. Since we have computed the derivatives of the functions $y = \sin x$ and $y = \cos x$ and since

$$\tan x = \frac{\sin x}{\cos x}, \quad \cot x = \frac{\cos x}{\sin x},$$

we can use Theorem 5.3 (to be more precise, the formula representing the derivative of the quotient, i.e. the third formula (5.16)) to compute the derivatives of the functions $y = \tan x$ and $y = \cot x$.

We shall see that everywhere except the points at which $\cos x = 0$

$$(\tan x)' = \frac{(\sin x)' \cos x - (\cos x)' \sin x}{\cos^2 x} = \frac{1}{\cos^2 x}.$$

Thus, we have

$$(\tan x)' = \frac{1}{\cos^2 x} = 1 + \tan^2 x$$

(for all the values of x except $x = \frac{\pi}{2} + \pi n$, where $n = 0, \pm 1, \ldots$). By analogy, everywhere except the points at which $\sin x = 0$,

$$(\cot x)' = \frac{(\cos x)' \sin x - (\sin x)' \cos x}{\sin^2 x} = -\frac{1}{\sin^2 x}.$$

Thus we have

$$(\cot x)' = -\frac{1}{\sin^2 x} = -(1 + \cot^2 x)$$

(for all the values of x except $x = \pi n$, where $n = 0, \pm 1, \ldots$).

5.4.5. The derivative of the function $y = \log ax$ $(0 < a \neq 1)$. Having taken as x any point of the half-line $x > 0$, and assuming that $|\Delta x| < x$, we can write

$$\Delta y = \log_a (x + \Delta x) - \log_a x = \log_a \frac{x + \Delta x}{x} = \log_a \left(1 + \frac{\Delta x}{x}\right).$$

Thus we have, for $\Delta x \neq 0$,

$$\frac{\Delta y}{\Delta x} = \frac{1}{\Delta x} \log_a \left(1 + \frac{\Delta x}{x}\right) = \log_a \left(1 + \frac{\Delta x}{x}\right)^{1/\Delta x} =$$

$$= \frac{1}{x} \log_a \left[\left(1 + \frac{\Delta x}{x}\right)^{x/\Delta x}\right]. \tag{5.26}$$

By virtue of the main result obtained in 4.6.3, the expression in brackets possesses, for $\Delta x \to 0$ (and for any fixed value of x), a limiting value equal to e. Then because of the continuity of the function $y = \log_a x$ at point $x = e$, there exists a limiting value (for $\Delta x \to 0$) on the right-hand side of (5.26) equal to $\frac{1}{x} \log_a e$. By the definition of the derivative, the indicated limiting value is equal to the derivative of the function $y = \log_a x$, i. e.

$$(\log_a x)' = \frac{1}{x} \log_a e$$

(for all the values of x belonging to the half-line $x > 0$). In the particular case $a = e$, we have

$$(\ln x)' = \frac{1}{x}.$$

5.5. THEOREM ON THE DERIVATIVE OF AN INVERSE FUNCTION

Theorem 5.4. *Let us postulate that the function $y = f(x)$ increases (or decreases) in some neighbourhood of a point x_0 and is continuous. Assume, in addition, that the function $y = f(x)$ is differentiable at point x_0 and the derivative $f'(x_0)$ is nonzero. Then there exists an inverse function $x = f^{-1}(y)$ which is defined in some neighbourhood of the corresponding point $y_0 = f(x_0)$, is differentiable at that point and possesses at that point a derivative equal to $\frac{1}{f'(x_0)}$.*

Proof. We shall note, first of all, that in the neighbourhood of point x_0 all the conditions of the corollary of Lemma 1 from 4.4.1, are satisfied for the function $y = f(x)$. In accordance with that corollary, there exists an inverse function $x = f^{-1}(y)$ defined in some neighbourhood of point $x_0 = f(x_0)$ and continuous in that neighbourhood. Let us give an arbitrary *nonzero* increment Δy to the argument y of this inverse function at point y_0. To this increment there corresponds an increment Δx of the inverse function, with $\Delta x \neq 0$ because of the increase (or decrease) of the function. Therefore, we are rightful in writing the following identity:

$$\frac{\Delta x}{\Delta y} = \frac{1}{\frac{\Delta y}{\Delta x}}. \tag{5.27}$$

Now let $\Delta y \to 0$ in identity (5.27). Then by virtue of the continuity of the inverse function $x = f^{-1}(y)$ at point y_0 and because of the difference form of the condition of continuity, $\Delta x \to 0$, as well.

But as Δx tends to zero, the denominator of the fraction on the right-hand side of (5.27) possesses, by the definition of a derivative, a limiting value equal to $f'(x_0) \neq 0$. Consequently, the right-hand side of (5.27) possesses, for $\Delta y \to 0$, a limiting

value equal to $\frac{1}{f'(x_0)}$. But then, the left-hand side of (5.27) possesses, for $\Delta y \to 0$, a limiting value too. In accordance with the definition of a derivative, the indicated limit is equal to* $\{f^{-1}(y_0)\}'$. Thus we have proved the differentiability of the inverse function at point y_0 and have obtained for its derivative the relation

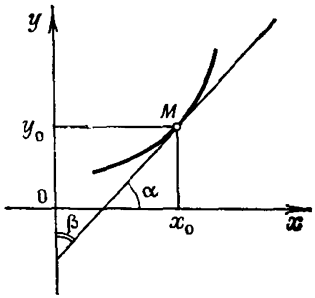


Fig. 5.4

$$\{f^{-1}(y_0)\}' = \frac{1}{f'(x_0)}. \tag{5.28}$$

The proof of Theorem 5.4 is complete.

This theorem has a simple geometrical meaning. Let us consider the graph of the function $y = f(x)$ (or of the inverse function) in the neighbourhood of point x_0. Let us assume that a point M on the graph corresponds to that point x_0 (Fig. 5.4). Then, evidently, the derivative $f'(x_0)$ is equal to the slope α of the tangent line, passing through point M, to the x-axis. The derivative of the inverse function $\{f^{-1}(y_0)\}'$ is equal to the slope β of the same tangent to the y-axis. Since the sum of the angles α and β is $\pi/2$, formula (5.28) expresses the obvious fact

$$\tan \beta = \frac{1}{\tan \alpha}.$$

5.6. COMPUTING THE DERIVATIVES OF AN EXPONENTIAL FUNCTION AND OF INVERSE TRIGONOMETRIC FUNCTIONS

Proceeding from Theorem 5.4 proved above, we shall continue calculating the derivatives of basic elementary functions.

5.6.1. The derivative of the exponential function $y = a^x$ $(0 < a \neq 1)$. Having been defined on the infinite interval, the exponential function $y = a^x$ serves as the inverse of the logarithmic function $x = \log_a y$ defined on the half-line $y > 0$. Since all the conditions of Theorem 5.4 are satisfied for the logarithmic function in the neighbourhood of any point y of the half-line $y > 0$, the function $y = a^x$ is differentiable at any point $x = \log_a y$ in accordance with this theorem, and formula

$$(a^x)' = \frac{1}{(\log_a y)'} = \frac{1}{\frac{1}{y}\log_a e} = \frac{y}{\log_a e}$$

is valid for its derivative. Making use of this formula and the relation $\log_a b = (1/\log_b a)$ known from elementary mathematics, and,

* The symbol $\{f^{-1}(y_0)\}'$ is used to denote the derivative of the inverse function at point y_0.

besides, taking into account that $y = a^x$, we finally obtain

$$(a^x)' = a^x \ln a.$$

This formula is valid for all the points x of the infinite interval. In the particular case $a = e$ the formula takes the form

$$(e^x)' = e^x.$$

5.6.2. The derivatives of inverse trigonometric functions. We shall begin with computing the derivative of the function $y = \arcsin x$. Having been defined on the interval $-1 < x < +1$, this function serves as the inverse of the function $x = \sin y$ defined on the interval $-(\pi/2) < y < +(\pi/2)$. Since all the conditions of Theorem 5.4 are complied with for the function $x = \sin y$ in the neighbourhood of any point y of the interval $-(\pi/2) < y < (\pi/2)$, the function $y = \arcsin x$ is differentiable at any point $x = \sin y$ in accordance with this theorem, and the formula

$$(\arcsin x)' = \frac{1}{(\sin y)'} = \frac{1}{\cos y} = \frac{1}{\sqrt{1-\sin^2 y}}. \tag{5.29}$$

is valid for its derivative. We have put the plus sign before the radical because $\cos y$ is positive everywhere on the interval $-\frac{\pi}{2} < y < \frac{\pi}{2}$. Taking into account that $\sin y = x$, we shall finally obtain, from formula (5.29)

$$(\arcsin x)' = \frac{1}{\sqrt{1-x^2}}.$$

As was mentioned in the process of derivation of this formula, it is valid for all x from the interval $-1 < x < +1$. The derivative of the function $y = \arccos x$ can be obtained by a complete analogy. Having been defined on the interval $-1 < x < +1$, this function serves as the inverse of the function $x = \cos y$ defined on the interval $0 < y < \pi$. Since all the conditions of Theorem 5.4 are satisfied for the function $x = \cos y$ in the neighbourhood of any point y of the interval $0 < y < \pi$, the function $y = \arccos x$ is differentiable at any point $x = \cos y$ in accordance with this theorem, and the formula

$$(\arccos x)' = \frac{1}{(\cos y)'} = -\frac{1}{\sin y} = -\frac{1}{\sqrt{1-\cos^2 y}}. \tag{5.30}$$

is valid for its derivative. We have taken into account that $\sin y = +\sqrt{1-\cos^2 y}$ since $\sin y$ is positive everywhere on the interval $0 < y < \pi$. Paying attention to the fact that $\cos y = x$, we shall

finally obtain, from formula (5.30),

$$(\arccos x)' = -\frac{1}{\sqrt{1-x^2}}.$$

As was noted in the process of its derivation, the formula obtained is valid for all the values of x from the interval $-1 < x < 1$.

Let us now compute the derivative of $y = \arctan x$. Having been defined on the infinite interval $-\infty < x < +\infty$, this function serves as the inverse of the function $x = \tan y$ defined on the interval $-(\pi/2) < y < (\pi/2)$. Since all the conditions of Theorem 5.4 are satisfied for the function $x = \tan y$ in the neighbourhood of any point y of the interval $-(\pi/2) < y < (\pi/2)$, the function $y = \arctan x$ is differentiable at any point $x = \tan y$ in accordance with this theorem, and the formula

$$(\arctan x)' = \frac{1}{(\tan y)'} = \frac{1}{1+\tan^2 y}.$$

is valid for its derivative. Taking into account that $\tan y = x$, we shall finally obtain

$$(\arctan x)' = \frac{1}{1+x^2}.$$

The formula obtained is valid for all the points x of the infinite interval.

It remains to calculate the derivative of the function $y = \operatorname{arc\,cot} x$. Having been defined on the infinite number line $-\infty < x < +\infty$, this function serves as the inverse of the function $x = \cot y$ defined on the interval $0 < y < \pi$. Since all the conditions of Theorem 5.4 are satisfied for the function $x = \cot y$ in the neighbourhood of any point y of the interval $0 < y < \pi$, the function $y = \operatorname{arc\,cot} x$ is differentiable at any point $x = \cot y$ in accordance with this theorem, and the formula

$$(\operatorname{arc\,cot} x)' = \frac{1}{(\cot y)'} = -\frac{1}{1+\cot^2 y}$$

is valid for its derivative. Taking into account that $\cot y = x$, we shall finally obtain

$$(\operatorname{arc\,cot} x)' = -\frac{1}{1+x^2}.$$

This formula is valid for all the points x of the infinite interval.

Thus we have calculated the derivatives of all the basic elementary functions with the exception of the power function with any real exponent.

We shall postpone the calculation of this last function till 5.8 and will now substantiate the rule of differentiating a composite function.

5.7. THE RULE OF DIFFERENTIATING A COMPOSITE FUNCTION

The aim of this section is to establish a rule making it possible to find the derivative of the composite function $y = f[\varphi(t)]$ if the derivatives of its constituent functions $y = f(x)$ and $x = \varphi(t)$ are known.

Theorem 5.5. *Let the function $x = \varphi(t)$ be differentiable at some point t_0 and the function $y = f(x)$ be differentiable at a corresponding point $x_0 = \varphi(t_0)$. Then the composite function $f[\varphi(t)]$ is differentiable at the indicated point t_0, the following formula* being valid for the derivative of this function:*

$$\{f[\varphi(t_0)]\}' = f'(x_0)\,\varphi'(t_0). \tag{5.31}$$

Proof. Let us assign an arbitrary *nonzero* increment Δt to the argument t at point t_0. To this increment there corresponds the increment Δx of the function $x = \varphi(t)$. In its turn, the increment Δx is associated with the increment Δy of the function $y = f(x)$ at point x_0. Since the function $y = f(x)$ is assumed to be differentiable at point x_0, the increment of this function at point x_0 can be written in the form (see Sec. 2)

$$\Delta y = f'(x_0)\,\Delta x + \alpha \Delta x, \tag{5.32}$$

where

$$\lim_{\Delta x \to 0} \alpha = 0.$$

Dividing equality (5.32) by Δt, we obtain

$$\frac{\Delta y}{\Delta t} = f'(x_0)\,\frac{\Delta x}{\Delta t} + \alpha\,\frac{\Delta x}{\Delta t}. \tag{5.33}$$

Now let Δt tend to zero in equality (5.33). Since the differentiability of the function $x = \varphi(t)$ at point t_0 implies the continuity of that function at point t_0, Δx tends to zero (at Δt tending to zero) by virtue of the difference form of the condition of continuity. Therefore, we can assert that there exists a limiting value

$$\lim_{\Delta t \to 0} \alpha = 0. \tag{5.34}$$

Besides, the requirement of the differentiability of the function $x = \varphi(t)$ at point t_0 implies the existence of a limiting value

$$\lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t} = \varphi'(t_0). \tag{5.35}$$

The existence of the limits (5.34) and (5.35) ensures the existence of a limit (for $\Delta t \to 0$) of the entire right-hand side of (5.33) equal to

* The symbol $\{f[\varphi(t_0)]\}'$ is used to denote the derivative of the composite function $y = f[\varphi(t)]$ at point $t = t_0$.

$f'(x_0)\,\varphi'(t_0)$. But this means that there exists a limiting value (for $\Delta t \to 0$) of the left-hand side of (5.35) as well. By the definition of a derivative, this limiting value is equal to the derivative of the composite function $f[\varphi(t)]$ at point t_0. Thus we have proved the differentiability of the composite function at point t_0 and established formula (5.31).

That was the proof of Theorem 5.5.

Remark. We have considered the composite function $y = f(x)$, where $x = \varphi(t)$, that is we have taken x as an intermediate argument and t as the final argument. The notation can, certainly, be changed. It is sometimes more convenient to consider a composite function of the form $y = f(u)$, where $u = \varphi(x)$, i.e. to take x as the final argument and some variable u as an intermediate argument. For that function the differentiation formula (5.31) takes the form

$$y' = \{f[\varphi(x)]\}' = f'(u)\,\varphi'(x) \tag{5.36}$$

(we have omitted zeros in the corresponding designations of the values of the arguments x and u because they played an auxiliary role).

Given below are some examples of the application of the rule of differentiation of a composite function we have just proved.

1°. It is required to compute the derivative of the function $y = e^{\arctan x}$. We shall treat this function as a composite function of the form $y = e^u$, where $u = \arctan x$. Making use of formula (5.36), we obtain

$$y' = (e^u)'(\arctan x)' = e^u \frac{1}{1+x^2} = e^{\arctan x} \frac{1}{1+x^2}.$$

2°. It is required to compute the derivative of the function $y = 2^{x^2}$. We shall treat this function as a composite function $y = 2^u$, where $u = x^2$. Making use of formula (5.36), we obtain

$$y' = (2^u)'(x^2)' = (2^u \ln 2)\,2x = 2^{x^2+1} x \ln 2.$$

3°. When discussing the indicated two examples, we have written out separately the functions constituting the given composite function. This is not necessary, of course, and in practical applications the differentiation of a composite function is carried out directly, without separating the constituent functions. For instance,

$$y = \arcsin 75x; \quad y' = \frac{1}{\sqrt{1-(75x)^2}}(75x)' = \frac{75}{\sqrt{1-(75x)^2}}$$

$\left(\text{here } |x| < \frac{1}{75}\right)$.

4°. Theorem 5.5 and the rule formulated in it can be successively carried over to the case of a composite function which is a superposition of three and more functions.

Let us consider an example of such a function. Suppose it is required to compute the derivative of the function $y = 5^{\operatorname{arc\,cot}(x^8)}$. Applying successively the rule of differentiating a composite function, we obtain

$$y' = (5^{\operatorname{arc\,cot}(x^8)} \ln 5) \frac{(-1)}{1+x^{16}} 8x^7.$$

5.8. THE LOGARITHMIC DERIVATIVE. THE DERIVATIVE OF A POWER FUNCTION WITH ANY REAL EXPONENT. THE TABLE OF THE DERIVATIVES OF BASIC ELEMENTARY FUNCTIONS

5.8.1. The concept of the logarithmic derivative of a function. Let the function $y = f(x)$ be *positive* and differentiable at a given point x. Then there exists at that point $\ln y = \ln f(x)$. Considering $\ln f(x)$ to be a composite function of the argument x, we can calculate the derivative of that function at point x assuming $y = f(x)$ to be an intermediate argument. Then we obtain

$$[\ln f(x)]' = \frac{y'}{y}. \tag{5.37}$$

The quantity determined by means of formula (5.37) is known as the *logarithmic derivative* of the function $y = f(x)$ at the given point x. By way of an example we shall calculate the logarithmic derivative of the so-called power-exponential function $y = u(x)^{v(x)}$. We know from 4.7.2 that this function is defined and continuous for all the values of x for which $u(x)$ and $v(x)$ are continuous and $u(x)$ is positive. Now we shall make an additional requirement that $u(x)$ and $v(x)$ shall be differentiable for the values of x being considered. Then, since $\ln y = v(x) \ln u(x)$, the logarithmic derivative will be equal to

$$\frac{y'}{y} = [v(x) \ln u(x)]' = v'(x) \ln u(x) + v(x) \frac{u'(x)}{u(x)}. \tag{5.38}$$

Taking into account that $y = u(x)^{v(x)}$, we obtain from equality (5.38) the following formula for the derivative of the power-exponential function:

$$y' = u(x)^{v(x)} \left[v'(x) \ln u(x) + v(x) \frac{u'(x)}{u(x)} \right].$$

5.8.2. The derivative of a power function with any real exponent. Let us compute the derivative of the power function $y = x^\alpha$ with an arbitrary real exponent α. We shall compute the derivative of this function for those values of x for which the function is defined

with any α, namely, for the values of x belonging to the half-line* $x > 0$. Bearing in mind that the function $y = x^\alpha$ is *positive* throughout the half-line $x > 0$, we shall calculate the logarithmic derivative of that function. Since $\ln y = \alpha \ln x$, the logarithmic derivative is equal to

$$\frac{y'}{y} = [\alpha \ln x]' = \frac{\alpha}{x},$$

whence, taking into account that $y = x^\alpha$, we can obtain the formula for the derivative of the power function

$$(x^\alpha)' = \alpha x^{\alpha-1}.$$

Thus we have calculated the derivatives of all the basic elementary functions. Collecting all the computed derivatives, we shall receive the following table which was given earlier, in Chap. 1.

5.8.3. The table of derivatives of basic elementary functions.

1°. $(x^\alpha)' = \alpha x^{\alpha-1}$. In particular, $\left(\frac{1}{x}\right)' = -\frac{1}{x^2}$, $(\sqrt{x})' = \frac{1}{2\sqrt{x}}$.

2°. $(\log_a x)' = \frac{1}{x}\log_a e$ $(x > 0,\ 0 < a \neq 1)$. In particular, $(\ln x)' = \frac{1}{x}$.

3°. $(a^x)' = a^x \ln a$ $(0 < a \neq 1)$. In particular, $(e^x)' = e^x$.

4°. $(\sin x)' = \cos x$.

5°. $(\cos x)' = -\sin x$.

6°. $(\tan x)' = \frac{1}{\cos^2 x} = 1 + \tan^2 x$ $\left(x \neq \frac{\pi}{2} + \pi n,\ \text{where}\ n = 0, \pm 1, \ldots\right)$.

7°. $(\cot x)' = -\frac{1}{\sin^2 x} = -(1 + \cot^2 x)$ $(x \neq \pi n,\ \text{where}\ n = 0, \pm 1, \ldots)$.

8°. $(\arcsin x)' = \frac{1}{\sqrt{1-x^2}}$ $(-1 < x < 1)$.

9°. $(\arccos x)' = -\frac{1}{\sqrt{1-x^2}}$ $(-1 < x < 1)$.

10°. $(\arctan x)' = \frac{1}{1+x^2}$.

11°. $(\operatorname{arc\,cot} x)' = -\frac{1}{1+x^2}$.

---

* In the case where $\alpha = 1/m$, where m is an odd integer, the function $y = x^\alpha$ is defined *throughout the infinite interval*. However, in this case as well it is sufficient to compute the derivative of the indicated function only for the values $x > 0$ since the function is *odd* and hence it is easy to obtain its derivative for the values $x < 0$.

In 4.4. we introduced the hyperbolic functions $y = \sinh x$, $y = \cosh x$, $y = \tanh x$ and $y = \coth x$, which are simple combinations of exponential functions. The definitions of these functions directly yield the following expressions for their derivatives:

12°. $(\sinh x)' = \cosh x$.

13°. $(\cosh x)' = \sinh x$.

14°. $(\tanh x)' = \dfrac{1}{\cosh^2 x}$.

15°. $(\coth x)' = -\dfrac{1}{\sinh^2 x}$ $(x \neq 0)$.

This table, together with the rules of differentiating a sum, a difference, a product and a quotient (i.e. formulas (5.16)) and the rule of differentiating a composite function, constitutes the basis of the differential calculus.

The rules and formulas of differentiation enable us to make a very important conclusion.

In 4.7 we introduced the concept of an *elementary function* which is a function expressible in terms of basic elementary functions by means of four arithmetic operations and superpositions successively applied a finite number of times. We can now assert that *the derivative of any elementary function is also an elementary function.* Thus, *the operation of differentiation does not take us outside the class of elementary functions.*

5.9. INVARIANCE OF THE FORM OF THE FIRST DIFFERENTIAL. SOME APPLICATIONS OF A DIFFERENTIAL

5.9.1. Invariance of the form of the first differential. We have established at the end of Sec. 2 that for the case when the argument x is an *independent variable* the differential of the function $y = f(x)$ can be determined from the formula

$$dy = f'(x)\,dx. \tag{5.39}$$

Here we shall prove that formula (5.39) is universal and is valid not only in the case where the argument x is an independent variable but also in the case of the argument x being itself a differentiable function of some new variable t. This property of the differential of a function is termed the *invariance of its form.*

Thus, suppose we are given a function $y = f(x)$, differentiable at some point x, whose argument x is a differentiable function $x = \varphi(t)$ of the argument t. In this case we can treat y *as a composite function* $y = f[\varphi(t)]$ of the argument t, and x as an intermediate argument. By virtue of Theorem 5.5 the derivative y with respect to t is determined from the formula

$$y'_t = f'(x)\,\varphi'(t). \tag{5.40}$$

Since we can treat t as an independent variable, the derivatives of the functions $x = \varphi(t)$ and $y = f[\varphi(t)]$ with respect to the argument t (in accordance with the law established at the end of 5.2) are equal to the ratio between the differentials of those functions and dt, i.e.

$$\varphi'(t) = \frac{dx}{dt}, \quad y' = \{f[\varphi(t)]\}' = \frac{dy}{dt}.$$

Substituting these values of the derivatives in formula (5.40), we make the formula look as follows:

$$\frac{dy}{dt} = f'(x)\frac{dx}{dt}. \tag{5.41}$$

Multiplying both sides of equality (5.41) by dt, we shall obtain for dy expression (5.39). Thus we have proved the invariance of the form of the first differential of the function, that is inferred that *both in the case of the argument x being an independent variable and in the case where the argument x is itself a differentiable function of some other variable, the differential dy of the function $y = f(x)$ is equal to the derivative of that function multiplied by the differential of the argument dx.*

The property of the invariance of a differential can be also formulated as follows: *the derivative of the function $y = f(x)$ is always* equal to the ratio between the differential of that function dy and the differential of the argument dx*, i.e.

$$f'(x) = \frac{dy}{dx}. \tag{5.42}$$

Equality (5.42) we have proved enables us from now on to use the ratio (dy/dx) to designate the derivative of the function $y = f(x)$ with respect to the argument x.

Note in conclusion that when equality (5.42) has been proved, the rule of differentiating a composite function takes on the form of a simple identity:

$$\frac{dy}{dt} = \frac{dy}{dx}\frac{dx}{dt}. \tag{5.43}$$

As simple a form is assumed by the rule of differentiating an inverse function:

$$\frac{dy}{dx} = \frac{1}{\frac{dx}{dy}}. \tag{5.44}$$

We want to emphasize, however, that equalities (5.43) and (5.44) cannot be considered as some new methods of proving Theorems 5.5

* That is, both in the case when the argument x is an independent variable and in the case where x is itself a differentiable function of some other variable.

and 5.4, for formulas (5.43) and (5.44) make essential use of the fact of the invariance of the first differential which we established precisely by means of Theorem 5.5.

5.9.2. Formulas and rules for computing differentials. We have proved that the differential dy of the function $y = f(x)$ is always equal to the derivative $f'(x)$ of that function multiplied by the differential of the argument dx. Thus, the table of the derivatives written out in 5.8.3 reduces to the corresponding table of differentials.

1°. $d(x^\alpha) = \alpha x^{\alpha-1}\,dx$. In a particular case $d\left(\frac{1}{x}\right) = -\frac{dx}{x^2}$, $d(\sqrt{x}) = \frac{dx}{2\sqrt{x}}$.

2°. $d(\log_a x) = \frac{\log_a e}{x}\,dx$ $(x>0,\ 0<a\neq 1)$. In a particular case $d(\ln x) = \frac{dx}{x}$.

3°. $d(a^x) = a^x \ln a\,dx$ $(0<a\neq 1)$. In a particular case, $d(e^x) = e^x\,dx$.

4°. $d(\sin x) = \cos x\,dx$.

5°. $d(\cos x) = -\sin x\,dx$.

6°. $d(\tan x) = \frac{dx}{\cos^2 x} = (1+\tan^2 x)\,dx$ $\left(x\neq \frac{\pi}{2}+\pi n,\ \text{where } n=0, \pm 1, \ldots\right)$.

7°. $d(\cot x) = -\frac{dx}{\sin^2 x} = -(1+\cot^2 x)\,dx$ $(x\neq \pi n,\ \text{where } n=0, \pm 1, \ldots)$.

8°. $d(\arcsin x) = \frac{dx}{\sqrt{1-x^2}}$ $(-1<x<1)$.

9°. $d(\arccos x) = -\frac{dx}{\sqrt{1-x^2}}$ $(-1<x<1)$.

10°. $d(\arctan x) = \frac{dx}{1+x^2}$.

11°. $d(\operatorname{arccot} x) = -\frac{dx}{1+x^2}$.

Formulas (5.16) and relation (5.39) immediately yield the following rules for computing a sum, a difference, a product and a quotient:

$$d(u\pm v) = du \pm dv,\quad d(uv) = v\,du + u\,dv,\quad d\left(\frac{u}{v}\right) = \frac{v\,du - u\,dv}{v^2}.$$

5.9.3. Application of differentials for establishing approximation formulas. Although, as we saw in Sec. 2, the differential dy of the function $y = f(x)$ is not equal to the increment Δy of that function,

the approximate relation

$$\Delta y \approx dy \tag{5.45}$$

is valid with an accuracy to within an infinitesimal of the order of smallness higher than that of Δx. The relative* error of this equality becomes infinitesimal for sufficiently small Δx. Formula (5.45) allows an approximate substitution of the differential dy of the function $y = f(x)$ for its increment Δy. The advantage of the substitution consists in the fact that the differential dy depends on Δx linearly while the increment Δy is, generally speaking, a more complex function of Δx.

Bearing in mind that the increment Δy of the function is determined by formula (5.1) and the differential dy by formula (5.14), we shall impart to the approximate equality (5.45) the following form:

$$f(x + \Delta x) - f(x) \approx f'(x)\,\Delta x$$

or

$$f(x + \Delta x) \approx f(x) + f'(x)\,\Delta x. \tag{5.46}$$

By means of formula (5.46) the function f for the values of the argument close to x (i.e. for small Δx) can be approximately replaced by a linear function.

In particular, a number of approximation formulas already known from Chap. 4 can be obtained from formula (5.46). Thus, assuming $f(x) = (1 + x)^{1/n}$, $x = 0$, we obtain

$$(1 + \Delta x)^{1/n} \approx 1 + \frac{\Delta x}{n}. \tag{5.47}$$

Putting $f(x) = \sin x$, $x = 0$, we obtain

$$\sin \Delta x \approx \Delta x. \tag{5.48}$$

Putting $f(x) = e^x$, $x = 0$, we receive

$$e^{\Delta x} \approx 1 + \Delta x. \tag{5.49}$$

Putting $f(x) = \ln(1 + x)$, $x = 0$, we obtain

$$\ln(1 + \Delta x) \approx \Delta x. \tag{5.50}$$

Each of the equalities (5.47)-(5.50) is accurate to within an infinitesimal of the order of smallness higher than that of Δx.

Equalities (5.47)-(5.50) in the form of accurate estimates were established at the end of 4.7.

---

* The relative error of this equality is determined by the ratio $\frac{\Delta y - dy}{\Delta x}$. Note that by the definition of a differential $\Delta y - dy = o(\Delta x)$.

5.10. DERIVATIVES AND DIFFERENTIALS OF HIGHER ORDERS

5.10.1. The concept of a derivative of the *n*th order. As was pointed out in 1.2, the derivative $f'(x)$ of the function $y = f(x)$, defined and differentiable in the interval (a, b), is a *function also defined in the interval* (a, b). It may so happen that the function $f'(x)$ itself is differentiable at some point x belonging to the interval (a, b), that is has a derivative at that point. Then the indicated derivative is referred to as the *second derivative* (or a *derivative of the 2nd order*) of the function $y = f(x)$ at point x and is symbolized as $f^{(2)}(x)$ or $y^{(2)}(x)$*.

After the concept of the second derivative has been introduced, we can successively introduce the concept of the third derivative, then of the fourth derivative and so on. If we postulate that the concept of the $(n-1)$th derivative has been introduced and that the $(n-1)$th derivative is differentiable at some point x belonging to the interval (a, b), i.e. possesses a derivative at that point, the indicated derivative is referred to as the *nth derivative* (or a *derivative of the nth order*) of the function $y = f(x)$ at point x and is designated as $f^{(n)}(x)$ or $y^{(n)}(x)$.

Thus, reasoning by induction, we have introduced the concept of the nth derivative by passing from the first derivative to the next and so on. The relation specifying the nth derivative has the form

$$y^{(n)} = [y^{(n-1)}]'. \tag{5.51}$$

A function that possesses a finite derivative of the nth order on the given set $\{x\}$ is usually said to be n times differentiable on that set. Higher-order derivatives find extensive application in physics. Here we shall restrict the discussion to the mechanical meaning of the second derivative. If the function $y = f(x)$ describes the law of motion of a particle in a straight line, then, as is known, the first derivative $f'(x)$ gives the instantaneous velocity of the moving particle at time moment x. In that case, the second derivative $f^{(2)}(x)$ is equal to the *rate of change of the velocity*, that is, it is equal to the *acceleration* of the moving particle at time moment x.

It might be well to point out that the techniques of computing derivatives of higher orders presuppose the ability of calculating the *derivatives only of the first order*. By way of example, let us calculate the derivatives of the nth order of some basic elementary functions.

5.10.2. Derivatives of the *n*th order of some functions.

1°. Let us compute the nth derivative of the power function $y = x^\alpha$ ($x > 0$, α being any real number). Consecutive differentiation

* The second derivative of the function $y = f(x)$ can be also symbolized as $f''(x)$ or $y''(x)$.

gives

$$y' = \alpha x^{\alpha-1},\ y^{(2)} = \alpha(\alpha-1)x^{\alpha-2},$$
$$y^{(3)} = \alpha(\alpha-1)(\alpha-2)x^{\alpha-3}, \ldots$$

Hence follows the general rule:

$$(x^{\alpha})^{(n)} = \alpha(\alpha-1)(\alpha-2)\ldots(\alpha-n+1)x^{\alpha-n}.$$

The rigorous proof of this rule can be easily carried out by induction.

In the particular case $\alpha = m$, where m is a natural number, we obtain

$$(x^{(m)})^{(m)} = m!,\ (x^m)^{(n)} = 0 \text{ for } n > m.$$

Thus the nth derivative of the polynomial of the mth order, for $n > m$, is equal to zero*.

2°. Next we shall compute the nth derivative of the exponential function $y = a^x$ $(0 < a \neq 1)$. Successive differentiation gives

$$y' = a^x \ln a,\quad y^{(2)} = a^x \ln^2 a,\quad y^{(3)} = a^x \ln^3 a, \ldots$$

The general formula established by the induction method has the form

$$(a^x)^{(n)} = a^x \ln^n a.$$

In particular,

$$(e^x)^{(n)} = e^x.$$

3°. Let us calculate the nth derivative of the function $y = \sin x$. The first derivative of this function can be written in the form $y' = \cos x = \sin(x + \pi/2)$. Thus, *as a result of the differentiation of the function* $y = \sin x$, *a quantity* $\pi/2$ *is added to the argument of that function.* Hence we obtain the formula

$$(\sin x)^{(n)} = \sin\left(x + n\frac{\pi}{2}\right).$$

4°. The formula

$$(\cos x)^{(n)} = \cos\left(x + n\frac{\pi}{2}\right)$$

can be derived by a complete analogy.

5°. In conclusion we shall compute the nth derivative of the so called *linear-fractional function* $y = (ax + b)/(cx + d)$, where a, b, c and d are some constants. Differentiating this function succes-

---

* In this case we additionally use the following obvious formula: $[Au(x) + Bv(x)]^{(n)} = Au^{(n)}(x) + Bv^{(n)}(x)$, where A and B are constants.

sively, we shall have

$$y' = \frac{a(cx+d)-c(ax+b)}{(cx+d)^2} = (ad-bc)(cx+d)^{-2},$$

$$y^{(2)} = (ad-bc)(-2)(cx+d)^{-3}c,$$

$$y^{(3)} = (ad-bc)(-2)(-3)(cx+d)^{-4}c^2, \ldots$$

It is easy to establish the general law

$$y^{(n)} = \left(\frac{ax+b}{cx+d}\right)^{(n)} = (ad-bc)(-1)^{n-1}n!(cx+d)^{-(n+1)}c^{n-1},$$

which can be substantiated by the method of induction.

5.10.3. Leibniz' formula for the *n*th order derivative of the product of two functions. While the rule for computing the first derivative of the sum or difference of the two functions $(u \pm v)' = u' \pm v'$ established above can be easily carried over (by induction, for instance) to the case of the nth derivative $(u \pm v)^{(n)} = u^{(n)} \pm v^{(n)}$, considerable difficulties arise in the process of computation of the nth derivative of the product of the two functions uv.

The corresponding rule is known as the *Leibniz formula* and has the following form:

$$(uv)^{(n)} = u^{(n)}v + C_n^1 u^{(n-1)}v' + C_n^2 u^{(n-2)}v^{(2)} + \\ + C_n^3 u^{(n-3)}v^{(3)} + \ldots + uv^{(n)}. \qquad (5.52)$$

It is easy to see the law governing the construction of the right-hand side of Leibniz' formula (5.52): *it coincides with the binominal formula* $(u+v)^n$, *with the only exception that the derivatives of the corresponding orders replace the powers of* u *and* v. The similarity becomes even more complete if, instead of the functions u and v, we write $u^{(0)}$ and $v^{(0)}$ respectively (that is, if we regard the function as a derivative of order zero).

Let us prove Leibniz' formula by induction. For $n = 1$ the formula assumes the form $(uv)' = u'v + uv'$, which coincides with the rule for differentiating a product of two functions established earlier (5.3). Therefore, postulating the validity of formula (5.52) for some number n it is sufficient to prove its validity for the next number $n+1$. Thus, suppose that formula (5.52) holds true for some number n. Let us differentiate the formula and combine the summands on the right-hand side as it is shown below:

$$(uv)^{(n+1)} = u^{(n+1)}v + [C_n^0 u^{(n)}v' + C_n^1 u^{(n)}v'] + \\ + [C_n^1 u^{(n-1)}v^{(2)} + C_n^2 u^{(n-1)}v^{(2)}] + \\ + [C_n^2 u^{(n-2)}v^{(3)} + C_n^3 u^{(n-2)}v^{(3)}] + \ldots + uv^{(n+1)}. \qquad (5.53)$$

(We have made use of the fact that $1 = C_n^0$.) It is known from elementary mathematics that for any number k not exceeding n the

formula*

$$C_n^k + C_n^{k-1} = C_{n+1}^k$$

holds true. Employing this formula we can rewrite equality (5.53) as follows:

$$(uv)^{(n+1)} = u^{(n+1)}v + C_{n+1}^1 u^{(n)}v' + C_{n+1}^2 u^{(n-1)}v^{(2)} + \ldots + uv^{(n+1)}).$$

Thus we have proved the validity of formula (5.52) for the number $(n+1)$ and completed the derivation of Leibniz' formula.

Example 1. It is required to compute the nth derivative of the function $y = x^2 \cos x$. Making use of Leibniz' formula, we shall put in it $u = \cos x$, $v = x^2$. In that case, for any number k we have $u^{(k)} = \cos\left(x + k\frac{\pi}{2}\right)$, $v' = 2x$, $v^{(2)} = 2$, $v^{(3)} = v^{(4)} = \ldots = 0$. We obtain

$$y^{(n)} = x^2 \cos\left(x + n\frac{\pi}{2}\right) + 2nx \cos\left[x + (n-1)\frac{\pi}{2}\right] +$$

$$+ n(n-1)\cos\left[x + (n-2)\frac{\pi}{2}\right].$$

Example 2. Let us now calculate the nth derivative of the function $y = x^3 e^x$. We shall again make use of Leibniz' formula, putting in it $u = e^x$, $v = x^3$. Then, for any number k we shall have $u^{(k)} = e^x$, $v' = 3x^2$, $v^{(2)} = 6x$, $v^{(3)} = 6$, $v^{(4)} = v^{(5)} = \ldots = 0$. We obtain

$$y^{(n)} = (x^3 + 3nx^2 + 3n(n-1)x + n(n-1)(n-2))e^x.$$

The examples we have considered show that Leibniz' formula is especially efficient in the case where one of the two functions being multiplied *possesses only a finite number of nonzero derivatives.*

5.10.4. Higher-order differentials. In this subsection we shall use the symbol δ, besides the symbol d, to designate a differential (i.e. we shall use the notation δx and δy instead of dx and dy where it is more convenient).

Suppose the function $y = f(x)$ is differentiable in some neighbourhood of a point x_0. Then the first differential dy of that function has the form** $dy = f'(x)\,dx$ and is the function of two variables, the point x and the quantity dx.

Assume, in addition, that the function $f'(x)$ is also differentiable at point x_0 and that the quantity dx has one and the same fixed value for all the points x of the neighbourhood of point x_0 under consideration.

---

* By the way, this formula can be easily verified.

** See 5.9.1, formula (5.39).

Under these assumptions there is a differential of the function $dy = f'(x)\,dx$ at point x_0 which we shall denote as $\delta(dy)$, this last differential being determined by the formula

$$\delta(dy) = \delta[f'(x)\,dx]\,|_{x=x_0} = [f'(x)\,dx]'_{x=x_0}\,\delta x = f''(x_0)\,dx\,\delta x. \quad (5.54)$$

Definition. *The value* $\delta(dy)$ *of the differential of the first differential* dy, *taken for* $\delta x = dx$, *is called the* **second differential** *of the function* $y = f(x)$ *(at point* x_0*) and is designated as* d^2y.

Formula (5.54) and the definition of the second differential yield

$$d^2y = f''(x_0)\,(dx)^2. \quad (5.55)$$

Make note of the fact that since we consider the quantity dx to be fixed, it immediately follows from the definition of the second differential that the second differential of the independent variable d^2x is equal to zero.

Differentials of higher orders can be determined by a complete analogy. Assuming that the derivative of the order $(n-1)$ of the function $y = f(x)$ is differentiable at point x_0 (that is, assuming that the function $y = f(x)$ possesses a derivative of the nth order at point x_0), we shall determine the *differential of the nth order* d^ny of the function $y = f(x)$ (at point x_0) as the differential $\delta(d^{n-1}y)$ of the differential of the $(n-1)$th order $d^{n-1}y$ taken for $\delta x = dx$.

The method of induction is the easiest method to derive the formula

$$d^ny = f^{(n)}(x_0)\,(dx)^n \quad (5.56)$$

for the differential of the nth order d^ny.

Indeed, formula (5.56) is valid for $n = 1$ and $n = 2$. Suppose that this formula is valid for some number $(n-1)$, i.e. assume that $d^{(n-1)}y = f^{(n-1)}(x)\,(dx)^{n-1}$.

Then, according to the definition of d^ny, we shall obtain*

$$d^ny = \delta(d^{n-1}y)\,|_{\delta x=dx} = \delta[f^{(n-1)}(x)\,(dx)^{n-1}]\,|_{\delta x=dx} =$$

$$= f^{(n)}(x)\,(dx)^{n-1}\,\delta x\,|_{\delta x=dx} = f^{(n)}(x)\,(dx)^n,$$

that is establish the validity of formula (5.56).

Formula (5.56) yields the following expression for the derivative of order n:

$$f^{(n)}(x) = \frac{d^ny}{(dx)^n}. \quad (5.56')$$

It is very important to point out that for $n > 1$ formulas (5.56) and (5.56′) are valid, in general, only in the case where x is an *independent variable* (i.e. the second and all the following differen-

* We neglect the index 0 of point x.

tials do not possess, in a general case, the property of the invariance of the form).

To make sure of this fact, let us discuss the problem of computing the second differential of the (twice differentiable) function $y = f(x)$ under the assumption that the variable x is the twice differentiable function of a certain argument t. Making use of equality (5.39) and the formula $\delta(uv) = v\delta u + u\delta v$, we obtain

$$d^2y = \delta(dy)\,|_{\delta x = dx} = \delta[f'(x)\,dx]\,|_{\delta x = dx} =$$
$$= \{dx\,\delta[f'(x)] + f'(x)\,\delta(dx)\}\,|_{\delta x = dx} =$$
$$= [dx \cdot f''(x)\,\delta(x)]\,|_{\delta x = dx} + f'(x)\,d^2x.$$

Thus we have $d^2y = f''(x)\,(dx)^2 + f'(x)\,d^2x$.

This last formula differs from (5.55) by the additional and, in general, nonzero member $f'(x)\,d^2x$.

5.11. DIFFERENTIATION OF A FUNCTION REPRESENTED PARAMETRICALLY

In this section we shall deal with the methods of calculating the derivatives of a function represented parametrically.

Let x and y be given as the functions of some parameter t: $x = \varphi(t)$, $y = \psi(t)$. Besides, we shall assume that the functions $\varphi(t)$ and $\psi(t)$ have the necessary number of derivatives with respect to the variable t in the range of this variable. We shall postulate, in addition, that in the neighbourhood of the considered point the function $x = \varphi(t)$ has an inverse function $t = \varphi^{-1}(x)$*. This postulation makes it possible to regard y as a function of the argument x.

Let us solve the problem of computing the derivatives y with respect to the argument x. We shall symbolize these derivatives as

$$y'_x,\ y^{(2)}_{x^2},\ y^{(3)}_{x^3},\ \ldots$$

By virtue of the invariance of the first differential we can write**

$$y'_x = \frac{dy}{dx},\quad dy = \psi'(t)\,dt,\quad dx = \varphi'(t)\,dt. \tag{5.57}$$

These formulas yield the following expression for the first derivative:

$$y'_x = \frac{\psi'(t)}{\varphi'(t)}. \tag{5.58}$$

The derivatives of the higher orders can be computed in a similar way. Thus, to compute the second derivative $y^{(2)}_{x^2}$, it is sufficient

* This is ensured by the existence of the first derivative $\varphi'(t)$, different from zero, in some neighbourhood of the point t being considered (see 15.2.4).

** In this case we take dy and dx at one and the same point t and for the same dt.

to represent it as

$$y''_{x^2} = \frac{d(y'_x)}{dx}$$

and make use of formula (5.58), the third formula (5.57) and the rule for differentiating a quotient.

Example. It is required to compute the first and the second derivative of a function represented parametrically:

$$\begin{cases} x = a(t - \sin t), \\ y = a(1 - \cos t), \quad -\infty < t < \infty. \end{cases}$$

The curve specified by these equations is called a *cycloid**. We obtain

$$y'_x = \frac{a \sin t}{a(1 - \cos t)} = \cot \frac{t}{2} \quad (t \neq 2\pi k, \text{ where } k \text{ is an integer}).$$

$$y''_{x^2} = \frac{\left[\cot \frac{t}{2}\right]'}{a(1 - \cos t)} = \frac{-1}{4a \sin^4 \frac{t}{2}}.$$

* A cycloid is a path of some fixed point of the circumference of a circle rolling along a straight line without slipping.

CHAPTER 6

INDEFINITE INTEGRAL

In this chapter we shall consider the problem of determining a function from the known derivative of that function. The urgency of this problem was elucidated in Chap. 1.

6.1. THE CONCEPTS OF AN ANTIDERIVATIVE OF A FUNCTION AND OF THE INDEFINITE INTEGRAL

6.1.1. The concept of an antiderivative of a function. Among the most important problems of mechanics is that of establishing the law of motion of a particle when its velocity is known, and the problem of determining the law of motion and the velocity of a particle from its acceleration*.

These problems lead to the mathematical problem of *finding a function from the given derivative of that function.*

Let us consider this problem.

Definition. *The function $F(x)$ is called an* **antiderivative** *(or a* **primitive function***) of the function $f(x)$ in the interval (a, b) if at any point x belonging to the interval (a, b) the function $F(x)$ is differentiable and possesses the derivative $F'(x)$ equal to $f(x)$.*

Remark. The definition of the antiderivative of the function $f(x)$ *on an infinite interval and an open half-line* is quite analogous**.

Examples. (1) The function $F(x) = \sqrt{1 - x^2}$ is the antiderivative of the function $f(x) = -\frac{x}{\sqrt{1-x^2}}$ on the interval $(-1, +1)$, for at any point x belonging to this interval $(\sqrt{1 - x^2})' = -\frac{x}{\sqrt{1-x^2}}$.

(2) The function $F(x) = \sin x$ is the antiderivative of the function $f(x) = \cos x$ on the infinite line $(-\infty, \infty)$, for at every point x of the infinite line $(\sin x)' = \cos x$.

---

* Instead of the acceleration of a particle we can preassign the force acting on the particle (for, in accordance with Newton's second law, the force determines the acceleration of this particle).

** And, generally, on any set $\{x\}$ *dense in itself*. For the definition of a set dense in itself see 2.3.

(3) The function $F(x) = \ln x$ is the antiderivative of the function $f(x) = (1/x)$ on the open half-line $x > 0$, for at every point x of this half-line $(\ln x)' = (1/x)$.

If $F(x)$ is the antiderivative of the function $f(x)$ on the interval (a, b), then, evidently, the function $F(x) + C$, where C is an arbitrary constant, is also an antiderivative of the function $f(x)$ on the interval (a, b).

The question naturally arises as to the connection between the antiderivatives of one and the same function $f(x)$. The following fundamental theorem is valid.

Theorem 6.1. *If $F_1(x)$ and $F_2(x)$ are arbitrary antiderivatives of the function $f(x)$ on the interval (a, b), then $F_1(x) - F_2(x) = C$ everywhere on that interval, where C is some constant.*

In other words, any two antiderivatives of the same function can differ only by a constant.

Proof. Let us put $\Phi(x) = F_1(x) - F_2(x)$. Since $F_1(x)$ and $F_2(x)$ are both differentiable on the interval (a, b), the function $\Phi(x)$ is also differentiable on the interval (a, b) by virtue of Theorem 5.3, with $\Phi'(x) = F_1'(x) - F_2'(x) = f(x) - f(x) = 0$ everywhere on that interval.

In 8.10 we shall prove. without using the results of this chapter*, Theorem 8.13 reading as follows: If the function $\Phi(x)$ is differentiable everywhere on the interval (a, b) and if $\Phi'(x) = 0$ everywhere on that interval, then the function $\Phi(x)$ is constant on the interval (a, b).

From this theorem we shall obtain $\Phi(x) = F_1(x) - F_2(x) = C = \text{const}$, and that is what we wanted to prove.

Corollary. If $F(x)$ is one of the antiderivatives of the function $f(x)$ on the interval (a, b), then *any* antiderivative $\Phi(x)$ of the function $f(x)$ on the interval (a, b) has the form $\Phi(x) = F(x) + C$, where C is some constant.

6.1.2. Indefinite integral.

Definition. *The collection of all the antiderivatives of the given function $f(x)$ on the interval (a, b) is called the* **indefinite integral** *of the function $f(x)$ (on that interval) and is symbolized as*

$$\int f(x)\,dx. \tag{6.1}$$

In this notation the sign $\int$ is called the *integral sign*, the expression $f(x)\,dx$ is the *element of integration*, and the function $f(x)$ itself is the *integrand*.

* We should point out that Chaps. 6 and 7 can be studied after Chap. 8 without prejudice to the understanding of the book. The point is that we want to familiarize the reader with the methods of integration, given in Chaps. 6 and 7, at an early stage.

If $F(x)$ is one of the antiderivatives of the function $f(x)$ on the interval (a, b), then, in accordance with the corollary of Theorem 6.1,

$$\int f(x)\,dx = F(x) + C, \tag{6.2}$$

where C is any constant.

It should be emphasized that *if there exists an antiderivative (and, consequently, the indefinite integral) of the function $f(x)$ on the interval (a, b), then the integration element in formula (6.1) is the differential of any of these antiderivatives.* Indeed, let $F(x)$ be any antiderivative of the function $f(x)$ on the interval (a, b), that is, $F'(x) = f(x)$ for all x belonging to the interval (a, b). Then $f(x)\,dx = F'(x)\,dx = dF$.

Examples. (1) $\int \frac{-x}{\sqrt{1-x^2}}\,dx = \sqrt{1-x^2} + C$ on the interval $-1 < x < 1$, since the function $F(x) = \sqrt{1-x^2}$ is one of the antiderivatives of the function $f(x) = \frac{-x}{\sqrt{1-x^2}}$ on the indicated interval.

(2) $\int \cos x\,dx = \sin x + C$ throughout the infinite interval $-\infty < x < \infty$, since the function $F(x) = \sin x$ is one of the antiderivatives of the function $f(x) = \cos x$ on the infinite interval.

In this chapter we shall not discuss the problem of the *existence* of antiderivatives (or the indefinite integrals) of extensive classes of functions. We shall only mention the fact that in 10.7 it will be proved that *for any function $f(x)$ continuous on the interval (a, b) there exists, on that interval, an antiderivative of the function (and the indefinite integral).*

The operation of finding an antiderivative or the indefinite integral (of function $f(x)$) is usually called *integration* (of the function $f(x)$).

6.1.3. Basic properties of the indefinite integral. We should point out, first of all, the two properties following immediately from the definition of the indefinite integral:

1°. $d\int f(x)\,dx = f(x)\,dx$.

2°. $\int dF(x) = F(x) + C$.

Property 1° implies that the signs d and $\int$ can be cancelled out if the differentiation sign stands before the sign of integration.

Property 2° means that the signs $\int$ and d can also be cancelled out if the integral sign precedes the differential sign, but in this case an arbitrary constant C must be added to $F(x)$.

To establish Property 1°, it is sufficient to differentiate both sides of formula (6.2) and take into account that $dF(x) = F'(x)\,dx = f(x)\,dx$.

To ascertain Property 2°, it is sufficient to make use of the equality $dF(x) = f(x)\,dx$ on the left-hand side of (6.2).

The following two properties are usually called *linear properties* of the integral.

3°. $\int [f(x) \pm g(x)]\,dx = \int f(x)\,dx \pm \int g(x)\,dx$.

4°. $\int [Af(x)]\,dx = A\int f(x)\,dx$ $(A = \text{const})$.

We want to emphasize the fact that the equality in formulas 3° and 4° is conditional and should be understood as the equality of the right-hand and left-hand sides with an accuracy to within an arbitrary constant summand (this is evident since each of the integrals entering into formulas 3° and 4° is specified with an accuracy to within an arbitrary constant summand).

Since two antiderivatives of one and the same function can differ only by a constant, to prove Property 3° it is sufficient to show that if $F(x)$ is an antiderivative of $f(x)$ and $G(x)$ is an antiderivative of $g(x)$, then the function $[F(x) \pm G(x)]$ is an antiderivative of the function $f(x) \pm g(x)$. This last statement follows immediately from the fact that the derivative of the (algebraic) sum of the functions is equal to the sum of the derivatives of those functions, i.e. $[F(x) \pm G(x)]' = F'(x) \pm G'(x) = f(x) \pm g(x)$. Property 4° can be proved by analogy. In that case use is made of the equality $[AF(x)]' = AF'(x) = Af(x)$.

6.1.4. Table of the basic indefinite integrals. In Chap. 5 we have obtained the table of the derivatives of the basic elementary functions (see 5.8) constituting the computational apparatus of differential calculus. Each formula of the table establishing the fact that one or another function $F(x)$ possesses a derivative equal to $f(x)$ leads, by virtue of the definition of the indefinite integral, to the corresponding formula of the integral calculus

$$\int f(x)\,dx = F(x) + C.$$

Thus we arrive at the following table of the basic indefinite integrals:

1°. $\int 0 \cdot dx = C$.

2°. $\int 1 \cdot dx = x + C$.

3°. $\int x^{\alpha}\,dx = \frac{x^{\alpha+1}}{\alpha+1} + C \quad (\alpha \neq -1)$.

4°. $\int \frac{dx}{x} = \ln|x| + C \quad (x \neq 0)$.

5°. $\int a^x\,dx=\frac{a^x}{\ln a}+C \quad (0<a\neq 1)$, $\int e^x\,dx=e^x+C$.

6°. $\int \sin x\,dx=-\cos x+C$. 7°. $\int \cos x\,dx=\sin x+C$.

8°. $\int \frac{dx}{\cos^2 x}=\int(1+\tan^2 x)\,dx=\tan x+C$

$\left(x\neq\frac{\pi}{2}+\pi n, \text{ where } n=0, \pm 1, \ldots\right)$.

9°. $\int \frac{dx}{\sin^2 x}=\int(1+\cot^2 x)\,dx=-\cot x+C$

$(x\neq \pi n, \text{ where } n=0, \pm 1, \ldots)$.

10°. $\int \frac{dx}{\sqrt{1-x^2}}=\begin{cases}\arcsin x+C,\\ -\arccos x+C\end{cases} (-1<x<1)$.

11°. $\int \frac{dx}{1+x^2}=\begin{cases}\arctan x+C,\\ -\operatorname{arccot} x+C.\end{cases}$

12°. $\int \frac{dx}{\sqrt{x^2\pm 1}}=\ln|x+\sqrt{x^2\pm 1}|+C$ (in the case of the minus sign $|x|>1$).

13°. $\int \frac{dx}{1-x^2}=\frac{1}{2}\ln\left|\frac{1+x}{1-x}\right|+C \quad (|x|\neq 1)$.

These formulas can be supplemented by those for the corresponding hyperbolic functions:

14°. $\int \sinh x\,dx=\cosh x+C$.

15°. $\int \cosh x\,dx=\sinh x+C$.

16°. $\int \frac{dx}{\cosh^2 x}=\tanh x+C$.

17°. $\int \frac{dx}{\sinh^2 x}=-\coth x+C \quad (x\neq 0)$.

A remark concerning formulas 4, 12 and 13 is in order. Formula 4 is valid for any interval which does not contain the value $x=0$. Indeed, if $x>0$, we infer from formula $(\ln x)'=1/x$ that $\int \frac{dx}{x}=\ln x+C$, and if $x<0$, formula $[\ln(-x)]'=1/x$ yields $\int \frac{dx}{x}=\ln(-x)+C$. Thus formula 4 is justified for any $x\neq 0$.

Formulas 12 and 13 are remarkable in that they have no analogues among the formulas from the table of derivatives.

However, to verify formulas 12 and 13, it is sufficient to make sure that the derivatives of the expressions entering into the right-hand sides of those formulas coincide with the corresponding integrands.

Our aim now is to supplement the table of indefinite integrals by the main methods and techniques of integration. But first we shall make an important remark.

In 4.7 we have introduced the concept of *elementary function* and in 5.8.3 established the fact that the derivative of any elementary function is an elementary function as well. In other words, we ascertained that the *operation of differentiation does not take us beyond the class of elementary functions.*

We must point out at once that the operation of integration is quite another matter. It can be proved that the integrals of some elementary functions are no longer elementary functions. Here are some examples of such integrals:

1°. $\int e^{-x^2}\,dx$.

2°. $\int \cos(x^2)\,dx$.

3°. $\int \sin(x^2)\,dx$.

4°. $\int \frac{dx}{\ln x}$ $(0<x\neq 1)$.

5°. $\int \frac{\cos x}{x}\,dx$ $(x\neq 0)$.

6°. $\int \frac{\sin x}{x}\,dx$.

Each of these integrals *is a function which is not elementary.* The indicated functions not only exist*, but also play a significant role in various problems of physics. Thus, for instance, integral 1 called *Poisson's integral* or *error integral* is widely used in statistical physics, in the theory of heat conduction and diffusion, integrals 2 and 3 named after Fresnel (*Fresnel integrals*) are employed in optics. Frequently encountered in applications are also integrals 4-6, the first of which is called an *integral logarithm* and the other two are known as an *integral cosine* and *integral sine* respectively.

For all the new functions we have enumerated (Poisson's integral, Fresnel's integral, integral logarithm, sine and cosine) there exist tables and graphs.

Since they are very important in applications, these functions have been studied as thoroughly as the basic elementary functions. We think it the right place to point out that the concept of a basic elementary function is very conditional.

---

* We have already mentioned that in 10.7 we shall prove the existence of the indefinite integral of any continuous function. The existence of integrals 1-6 is ensured by the continuity of integrands.

6.2. MAIN METHODS OF INTEGRATION

6.2.1. Integration by a change of variable (substitution). A change of variable is one of the most efficient methods of integration. This method is based on the following elementary statement.

Let the function $t = \varphi(x)$ *be defined and differentiable on some set* $\{x\}$*, and let* $\{t\}$ *be the set of all values of that function. Assume now that the function* $g(t)$ *possesses on the set* $\{t\}$ *an antiderivative* $G(t)$*, i.e.*

$$\int g(t)\,dt = G(t) + C. \tag{6.3}$$

Then, everywhere on the set $\{x\}$ *there exists an antiderivative for the function* $g[\varphi(x)]\,\varphi'(x)$ *equal to* $G[\varphi(x)]$*, i.e.*

$$\int g[\varphi(x)]\,\varphi'(x)\,dx = G[\varphi(x)] + C. \tag{6.4}$$

To prove this assertion, it is sufficient to apply the rule for differentiating a composite function**

$$\frac{d}{dx}\{G[\varphi(x)]\} = G'[\varphi(x)]\,\varphi'(x)$$

and take into account that by the definition of an antiderivative $G'(t) = g(t)$. Now suppose that we have to compute the integral

$$\int f(x)\,dx. \tag{6.5}$$

In some cases, it is possible to take as a new variable such a differentiable function $t = \varphi(x)$ that the equality

$$f(x) = g[\varphi(x)]\,\varphi'(x) \tag{6.6}$$

holds true, and that the function $g(t)$ can be easily integrated, i.e. the integral

$$\int g(t)\,dt = G(t) + C$$

can be easily computed. The statement proved above enables us to write the following formula for the integral (6.5):

$$\int f(x)\,dx = G[\varphi(x)] + C. \tag{6.7}$$

This technique of computing integral (6.5) is known as the *integration by a change of variable.*

---

* This set is either an open interval or a closed interval, or a half-line, or else an infinite straight line.

** See 5.7.

Surely, this technique cannot be applied to every integral. Besides, it should be pointed out that the correct choice of substitution largely depends on the proficiency of the person carrying out the computations. Here are some examples illustrating the method just discussed.

1°. It is required to calculate $\int \cos 2x\, dx$. To perform the calculation, it is necessary to make the following simple substitution: $t = 2x$, $dt = 2dx$. As a result of the substitution we obtain

$$\int \cos 2x\, dx = \int \frac{1}{2} \cos t\, dt = \frac{1}{2} \sin t + C = \frac{1}{2} \sin 2x + C.$$

2°. It is required to calculate $\int \frac{dx}{x+a}$. This integral can be computed by means of the substitution $t = x + a$, $dt = dx$.

The substitution will yield

$$\int \frac{dx}{x+a} = \int \frac{dt}{t} = \ln |t| + C = \ln |x+a| + C \quad (x \neq -a).$$

3°. It is required to calculate $\int e^{\cos x} \sin x\, dx$. It is easy to see that this integral can be computed by means of the substitution $t = \cos x$. Indeed, in this case $dt = -\sin x\, dx$ and

$$\int e^{\cos x} \sin x\, dx = -\int e^t\, dt = -e^t + C = -e^{\cos x} + C.$$

4°. It is required to calculate $\int \frac{(\arctan x)^{100}}{1+x^2} dx$. To compute this integral, it is convenient to make the substitution $t = \arctan x$. Indeed, with this substitution $dt = \frac{dx}{1+x^2}$ and

$$\int \frac{(\arctan x)^{100}}{1+x^2} dx = \int t^{100}\, dt = \frac{t^{101}}{101} + C = \frac{(\arctan x)^{101}}{101} + C.$$

5°. It is required to calculate $\int (7x - 9)^{2\,999}\, dx$. This integral can certainly be reduced to the sum of three thousand tabulated integrals by writing out the integrand function with the aid of Newton's binomial formula. But it is much easier to make the substitution $t = 7x - 9$, $dt = 7dx$, as a result of which we obtain

$$\int (7x-9)^{2\,999}\, dx = \frac{1}{7} \int t^{2\,999}\, dt = \frac{t^{3\,000}}{21\,000} + C = \frac{(7x-9)^{3\,000}}{21\,000} + C.$$

6°. It is required to compute $\int \frac{dx}{\cos x}$. To choose the substitution which would enable us to find this integral, let us rewrite it

in the form

$$\int \frac{dx}{\cos x} = \int \frac{\cos x\, dx}{\cos^2 x} = \int \frac{\cos x\, dx}{1-\sin^2 x}.$$

Now it is clear that we must put $t = \sin x$, $dt = \cos x\, dx$. As a result we obtain

$$\int \frac{dx}{\cos x} = \int \frac{dt}{1-t^2} = \frac{1}{2} \ln \left| \frac{1+t}{1-t} \right| + C = \ln \left| \tan \left(\frac{x}{2} + \frac{\pi}{4} \right) \right| + C.$$

7°. It is required to calculate $\int \frac{x^3\, dx}{(2x)^8+1}$. The following substitution is convenient: $t=(2x)^4$, $dt=64x^3\, dx$. In this case

$$\int \frac{x^3\, dx}{(2x)^8+1} = \frac{1}{64} \int \frac{dt}{t^2+1} = \frac{\arctan t}{64} + C = \frac{\arctan (2x)^4}{64} + C.$$

8°. It is required to calculate $\int \frac{dx}{(x^2+a^2)^{3/2}}$. To perform the calculation, it is convenient to make the trigonometric substitution $t = \arctan \frac{x}{a}$, $x = a \tan t$, $dx = a \frac{dt}{\cos^2 t}$.

As a result of this substitution the integral assumes the form

$$\int \frac{dx}{(x^2+a^2)^{3/2}} = \frac{1}{a^2} \int \cos t\, dt = \frac{\sin t}{a^2} + C = \frac{\tan t}{a^2 \sqrt{1+\tan^2 t}} + C =$$

$$= \frac{x}{a^2 \sqrt{x^2+a^2}} + C.$$

9°. It is required to calculate $\int \frac{dx}{(a^2-x^2)^{3/2}}$. Here it is convenient to make the substitution $t = \arcsin \frac{x}{a}$, $x = a \sin t$, $dx = a \cos t\, dt$. In this case

$$\int \frac{dx}{(a^2-x^2)^{3/2}} = \frac{1}{a^2} \int \frac{dt}{\cos^2 t} = \frac{\tan t}{a^2} + C =$$

$$= \frac{1}{a^2} \frac{\sin t}{\sqrt{1-\sin^2 t}} + C = \frac{x}{a^2 \sqrt{a^2-x^2}} + C.$$

10°. It is required to calculate $\int \sqrt{\frac{a+x}{a-x}}\, dx$. To compute this integral, it is convenient to make the substitution $2t = \arccos \frac{x}{a}$, $x = a \cos 2t$, $dx = -2a \sin 2t\, dt$. We obtain

$$\int \sqrt{\frac{a+x}{a-x}}\, dx = -4a \int \cos^2 t\, dt =$$

$$= -4a \int \left(\frac{1}{2} + \frac{1}{2} \cos 2t \right) dt = -2at - 2a \int \cos 2t\, dt =$$

$$= -2at - a \sin 2t + C = -a \left[\arccos \frac{x}{a} + \sqrt{1-\left(\frac{x}{a}\right)^2} \right] + C.$$

6.2.2. Integration by parts. *Integration by parts* is one of the most efficient methods of integration. It is based on the following assertion.

Assume that each of the functions $u(x)$ and $v(x)$ is differentiable on the set $\{x\}$ and, besides, there exists, on this set, an antiderivative of the function $v(x)\,u'(x)$. Then, on the set $\{x\}$, there exists an antiderivative of the function $u(x)\,v'(x)$ as well, the formula

$$\int u(x)\,v'(x)\,dx = u(x)\,v(x) - \int v(x)\,u'(x)\,dx \tag{6.8}$$

holding true.

Remark. The definition of the differential and the property of the invariance of its form permit writing formula (6.8) in the form

$$\int u\,dv = u(x)\,v(x) - \int v\,du. \tag{6.9}$$

To prove the formulated assertion, we shall write the formula for the derivative of the product of the two functions $u(x)$ and $v(x)$

$$[u(x)\,v(x)]' = u(x)\,v'(x) + u'(x)\,v(x). \tag{6.10}$$

Let us multiply equality (6.10) by dx and integrate both sides of the equality thus obtained. Since, by the hypothesis, $\int v(x)\,u'(x)\,dx$ and $\int [u(x)\,v(x)]'\,dx = u(x)\,v(x) + C$ for all x on the set $\{x\}$ (see Property 2°_1 in 1.3), there also exists an integral $\int u(x)\,v'(x)\,dx$ for all x belonging to the set $\{x\}$, formula (6.8) (or (6.9)) being valid in this case.

Formula (6.9) *reduces the problem of computing the integral* $\int u\,dv$ *to that of computing the integral* $\int v\,du$. In a number of concrete cases this integral can be computed quite easily.

The computation of the integral $\int u\,dv$ by means of formula (6.9) is referred to as *integration by parts*. Note that when formula (6.9) is applied to concrete cases it is very convenient to use the table of differentials given in 5.9.2.

Let us consider some examples.

1°. Calculate the integral $I = \int x^n \ln x\,dx$ $(n \neq -1)$. Assuming $u = \ln x$, $dv = x^n\,dx$ and using formula (6.9), we obtain $du = \frac{dx}{x}$, $v = \frac{x^{n+1}}{n+1}$,

$$I = \frac{x^{n+1}}{n+1}\ln x - \frac{1}{(n+1)}\int x^n\,dx = \frac{x^{n+1}}{n+1}\left(\ln x - \frac{1}{n+1}\right) + C.$$

2°. Calculate now the integral $I=\int x \arctan x\,dx$. Assuming $u=\arctan x$, $dv=x\,dx$ and using formula (6.9), we shall have $du=\frac{dx}{1+x^2}$, $v=\frac{x^2}{2}$,

$$I=\frac{x^2}{2}\arctan x-\frac{1}{2}\int\frac{x^2}{1+x^2}dx=\frac{x^2}{2}\arctan x-\frac{1}{2}\int\frac{[(1+x^2)-1]}{1+x^2}dx=$$

$$=\frac{x^2}{2}\arctan x-\frac{1}{2}\int dx+\frac{1}{2}\int\frac{dx}{1+x^2}=\frac{x^2+1}{2}\arctan x-\frac{x}{2}+C.$$

3°. Calculate the integral $I=\int x^2\cos x\,dx$. Let us first apply formula (6.9) assuming $u=x^2$, $dv=\cos x\,dx$. We shall obtain $du=2x\,dx$, $v=\sin x$, $I=x^2\sin x-2\int x\sin x\,dx$. To compute the last integral, let us apply formula (6.9) once again, this time putting $u=x$, $dv=\sin x\,dx$. We receive $du=dx$, $v=-\cos x$, $I=x^2\sin x+2x\cos x-2\int\cos x\,dx=(x^2-2)\sin x+2x\cos x+C$. Thus, we have calculated the integral $\int x^2\cos x\,dx$ by two-fold integration by parts. It is easy to realize that the integral $\int x^n\cos x\,dx$ (where n is an arbitrary positive integer) can be calculated in a similar way, by means of n-fold integration by parts.

4°. Calculate now the integral $I=\int e^{ax}\cos bx\,dx$ ($a=\text{const}$, $b=\text{const}$). Let us first apply formula (6.9) assuming $u=e^{ax}$, $dv=\cos bx\,dx$. We obtain $du=ae^{ax}\,dx$, $v=\frac{\sin bx}{b}$,

$$I=\frac{e^{ax}\sin bx}{b}-\frac{a}{b}\int e^{ax}\sin bx\,dx.$$

To compute this last integral, let us apply formula (6.9) once again, this time assuming $u=e^{ax}$, $dv=\sin bx\,dx$. We obtain $du=ae^{ax}\,dx$, $v=-\frac{\cos bx}{b}$,

$$I=\frac{e^{ax}\sin bx}{b}+\frac{a}{b^2}e^{ax}\cos bx-\frac{a^2}{b^2}I. \tag{6.11}$$

Thus, by means of two-fold integration of I by parts we shall obtain for the integral I the first-order equation (6.11). This equation yields

$$I=\frac{a\cos bx+b\sin bx}{a^2+b^2}e^{ax}.$$

It is known from experience, that a large part of the integrals obtained by means of integration by parts, can be classified into the following *three groups*:

(1) The *first* group contains the integrals whose integrand function includes, as a factor, one of the following functions: $\ln x$, $\arcsin x$, $\arccos x$, $\arctan x$, $(\arctan x)^2$, $(\arccos x)^2$, $\ln \varphi(x)$, ... (see Examples 1° and 2° discussed above). To compute the integrals of the first group, it is good practice to apply formula (6.9), assuming in it $u(x)$ to be equal to one of the indicated functions*.

(2) The *second* group includes the integrals of the form $\int (ax + b)^n \cos(cx)\,dx$, $\int (ax + b)^n \sin(cx)\,dx$, $\int (ax + b)^n e^{cx}\,dx$, where a, b and c are some constants and n is an arbitrary positive integer (see Example 3° discussed above). The integrals of the *second* group are obtained by means of the n-fold application of the formula for integration by parts (6.9); in this case $(ax + b)$ in the respective degree must be each time substituted for $u(x)$. After each integration by parts this degree becomes lower by unity.

(3) The *third* group includes the integrals of the form $\int e^{ax} \cos bx\,dx$, $\int e^{ax} \sin bx\,dx$, $\int \sin(\ln x)\,dx$, $\int \cos(\ln x)\,dx$, ... (see Example 4° considered above). Designating any of the integrals of this group by I and twice integrating by parts, we compile for I a first-order equation.

Of course, the indicated three groups do not include all the integrals, without exception, obtainable by means of integration by parts. Given below are examples of integrals not contained in any of the enumerated three groups but obtainable by means of formula (6.9).

5°. Let us compute the integral $I = \int \frac{x\,dx}{\cos^2 x}$. This integral is not contained in any of the groups enumerated above. Nonetheless, applying formula (6.9) and putting in it $u = x$, $dv = \frac{dx}{\cos^2 x}$ we obtain $du = dx$, $v = \tan x$,

$$I = x \tan x - \int \tan x\,dx = x \tan x - \int \frac{\sin x\,dx}{\cos x} =$$

$$= x \tan x + \int \frac{d(\cos x)}{\cos x} = x \tan x + \ln|\cos x| + C.$$

6°. Let us finally calculate the integtal $K_\lambda = \int \frac{dt}{(t^2 + a^2)^\lambda}$ which is very significant for our further aims. Here $a = \text{const}$, $\lambda = 1, 2, \ldots$ This integral does not enter into the three enumerated groups either. To compute this integral, we shall establish

* In the case where the integrand function contains as a factor $(\arctan x)^2$, $(\arccos x)^2$, ..., the formula for integration by parts (6.9) should be applied twice.

for it a recurrence formula which reduces the problem of computing K_λ to that of computing $K_{\lambda-1}$.

We can write (for $\lambda \neq 1$)

$$K_\lambda = \frac{1}{a^2}\int \frac{a^2\,dt}{(t^2+a^2)^\lambda} = \frac{1}{a^2}\int \frac{[(t^2+a^2)-t^2]\,dt}{(t^2+a^2)^\lambda} =$$

$$= \frac{1}{a^2}\int \frac{dt}{(t^2+a^2)^{\lambda-1}} - \frac{1}{2a^2}\int t\,\frac{2t\,dt}{(t^2+a^2)^\lambda} =$$

$$= \frac{1}{a^2}K_{\lambda-1} - \frac{1}{2a^2}\int t\,\frac{d(t^2+a^2)}{(t^2+a^2)^\lambda}.$$

To compute the last integral, we shall apply formula (6.9) for integration by parts assuming in it $u=t$, $dv=\frac{d(t^2+a^2)}{(t^2+a^2)^\lambda}$. We obtain $du=dt$, $v=\frac{1}{(\lambda-1)(t^2+a^2)^{\lambda-1}}$,

$$K_\lambda = \frac{1}{a^2}K_{\lambda-1} + \frac{t}{2a^2(\lambda-1)(t^2+a^2)^{\lambda-1}} - \frac{1}{2a^2(\lambda-1)}K_{\lambda-1}.$$

The last equality yields the recurrence formula

$$K_\lambda = \frac{t}{2a^2(\lambda-1)(t^2+a^2)^{\lambda-1}} + \frac{1}{a^2}\,\frac{(2\lambda-3)}{(2\lambda-2)}\,K_{\lambda-1}. \qquad (6.12)$$

Let us make sure that the recurrence formula (6.12) makes it possible to compute the integral K_λ for any $\lambda = 2, 3, \ldots$ Indeed it is easy to calculate the integral K_1:

$$K_1 = \int \frac{dt}{t^2+a^2} = \frac{1}{a}\int \frac{d\left(\frac{t}{a}\right)}{\left(\frac{t}{a}\right)^2+1} = \frac{1}{a}\arctan\frac{t}{a} + C.$$

After we have calculated the integral K_1, we put $\lambda = 2$ in formula (6.12) and easily compute the integral K_2. Then, knowing the value of K_2 and assuming $\lambda = 3$, in formula (6.12), we can easily calculate K_3. Proceeding in this way, we shall calculate the integral K_λ for any natural λ.

CHAPTER 7

COMPLEX NUMBERS. ALGEBRA OF POLYNOMIALS. INTEGRATION IN TERMS OF ELEMENTARY FUNCTIONS

It was pointed out in the previous chapter that the indefinite integral of an elementary function is not, in general, an elementary function. Nevertheless, there are wide classes of functions whose integrals are elementary functions. (Such classes of functions will be called *integrable in terms of elementary functions.*) The investigation of the indicated classes of functions is the main aim of the present chapter. Since the *class of rational functions* is one of the principal classes among the indicated ones, we must first of all elucidate the concepts of polynomials and rational functions. But this, in its turn, requires the elucidation of the concept of complex numbers.

7.1. BRIEF INFORMATION ON COMPLEX NUMBERS

Two real numbers x and y will be referred to as an ordered pair if it is indicated which of them is the first number and which is the second. The ordered pair of real numbers x and y will be designated as (x, y), the first place being occupied by the first member x of the pair.

A complex number is an ordered pair (x, y) of real numbers, the first of which x is called the real part and the second y is called the imaginary part of this complex number.

In the case when the imaginary part y is equal to zero, the corresponding pair $(x, 0)$ is agreed to be considered identical with the real number x. This agreement makes it possible to consider the set of all real numbers as a part of the set of complex numbers.

Two complex numbers $z_1 = (x_1, y_1)$ and $z_2 = (x_2, y_2)$ are said to be equal if $x_1 = x_2$ and $y_1 = y_2$. A complex number $z = (x, y)$ is assumed to be zero if $x = 0$ and $y = 0$.

Let us define the operations of addition and multiplication of complex numbers. Since real numbers constitute a part of the set of complex numbers, these operations should be defined so that when applied to two real numbers they would lead to the definitions of a sum and a product of real numbers already cited in 2.2.

The sum of two complex numbers $z_1 = (x_1, y_1)$ *and* $z_2 = (x_2, y_2)$ *will be a complex number* z *of the form*

$$z = (x_1 + x_2, y_1 + y_2). \tag{7.1}$$

The product of two complex numbers $z_1 = (x_1, y_1)$ *and* $z_2 = (x_2, y_2)$ *will be a complex number* z *of the form*

$$z = (x_1x_2 - y_1y_2, x_1y_2 + x_2y_1). \tag{7.2}$$

It is easy to verify that the sum and the product of complex numbers possess the same properties as the sum and the product of real numbers. Namely, the following properties hold true:

1°. $z_1 + z_2 = z_2 + z_1$ (commutativity of a sum).

2°. $(z_1 + z_2) + z_3 = z_1 + (z_2 + z_3)$ (associative property of a sum).

3°. $z + (0, 0) = z$ (a special role of the number (0,0)).

4°. For every number $z = (x, y)$ there exists an inverse number $z' = (-x, -y)$ such that $z + z' = (0, 0)$.

5°. $z_1 \cdot z_2 = z_2 \cdot z_1$ (commutativity of a product).

6°. $(z_1 \cdot z_2) \cdot z_3 = z_1 \cdot (z_2 \cdot z_3)$ (associative property of a product).

7°. $z \cdot (1, 0) = z$ (a special role of the number (1, 0)).

8°. For any nonzero complex number $z = (x, y)$ there exists an inverse number $\frac{1}{z} = \left(\frac{x}{x^2+y^2}, -\frac{y}{x^2+y^2}\right)$ such that $z \cdot \frac{1}{z} = (1, 0)$.

9°. $(z_1 + z_2) \cdot z_3 = z_1 \cdot z_3 + z_2 \cdot z_3$ (distributive property of a product with respect to a sum).

Properties 1°-9° enable us to assert that all the rules of elementary algebra referring to arithmetic operations and combinations of equalities hold true for complex numbers as well.

These properties also allow us to infer that *a subtraction of complex numbers* is an operation inverse of addition and *a division of complex numbers* is an operation inverse of multiplication.

The difference between two complex numbers $z_1 = (x_1, y_1)$ *and* $z_2 = (x_2, y_2)$ *is such a complex number* z *which, being added to* z_2, *produces the sum* z_1. With the aid of properties 1°-4° it is easy to establish the existence and the uniqueness of the difference between any two complex numbers*.

It is easy to verify that the difference between two complex numbers $z_1 = (x_1, y_1)$ and $z_2 = (x_2, y_2)$ is a complex number z of the form

$$z = (x_1 - x_2, y_1 - y_2). \tag{7.3}$$

The quotient of two complex numbers $z_1 = (x_1, y_1)$ *and* $z_2 = (x_2, y_2)$, *the second of which is nonzero, is such a complex number* z *which, being multiplied by* z_2, *gives* z_1. It can be established with the aid of properties 5°-8° that the only quotient of the two indicated complex

* This is done in quite the same way as for real numbers (see 2.2.3).

numbers is a complex number z of the form

$$z = \left(\frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2}, \frac{x_2y_1 - x_1y_2}{x_2^2 + y_2^2} \right). \tag{7.4}$$

A special role in operations with complex numbers is played by a number representable by the pair (0, 1) and symbolized by the letter i. Multiplying this pair by itself (i.e. squaring it), we obtain, by virtue of the definition of a product of complex numbers,

$(0, 1) \cdot (0, 1) = (-1, 0) = -1$, i.e. $i^2 = -1$.

Bearing this in mind, we can represent any complex number $z = (x, y)$ in the form

$$z = (x, y) = (x, 0) + (0, y) = (x, 0) + (y, 0) \cdot (0, 1) = x + iy.$$

In what follows we shall frequently use the representation $z = x + iy$ for the complex number $z = (x, y)$. This representation and the fact that the number i can be treated as a factor whose square is equal to -1 make it possible to carry out operations on complex numbers in just the same way as they are performed on algebraic polynomials.

The complex number $\bar{z} = (x, -y) = x - iy$ is usually called the *conjugate of the complex number* $z = (x, y) = x + iy$.

It is evident that *a complex number is equal to zero if and only if its conjugate number is equal to zero**.

The Cartesian rectangular system of coordinates is very convenient for geometrical representation of complex numbers. In this case the complex number $z = (x, y)$ is represented either by a point M with the coordinates (x, y) or a vector $\overrightarrow{OM}$ running from the origin to the point M.

With this method of representation, the addition and subtraction of complex numbers reduce to the addition and subtraction of the corresponding vectors (this is clear from formulas (7.1) and (7.3)).

If, alongside the Cartesian system of coordinates, we introduce a polar coordinate system so that the pole is at the origin O of the Cartesian coordinate system and the direction of the polar axis coincides with that of the positive x-axis, then, as is known, the Cartesian coordinates (x, y) and the polar coordinates (ρ, θ) of any point M can be related by the formulas:

$$\left.\begin{aligned} x &= \rho \cos \theta, \\ y &= \rho \sin \theta \end{aligned}\right\} \quad \begin{aligned} \rho &= \sqrt{x^2 + y^2}, \\ \theta &= \begin{cases} \arctan \dfrac{y}{x} & \text{for } x > 0, \\ \arctan \dfrac{y}{x} + \pi \operatorname{sgn} y & \text{for } x < 0, \\ \dfrac{\pi}{2} \operatorname{sgn} y & \text{for } x = 0. \end{cases} \end{aligned} \tag{7.5}$$

* Because the equalities $\left.\begin{aligned} x &= 0, \\ y &= 0 \end{aligned}\right\}$ and $\left.\begin{aligned} x &= 0, \\ -y &= 0 \end{aligned}\right\}$ are equivalent.

Formulas (7.5) lead to the *trigonometric form* of representation of the complex number $z = (x, y)$:

$$z = (x, y) = x + iy = (\rho \cos \theta, \rho \sin \theta) = \rho (\cos \theta + i \sin \theta). \tag{7.6}$$

In the trigonometric representation (7.6) the number ρ is called the *modulus* and the angle θ is called the *argument* (*amplitude*) of the complex number. The specification of the argument θ is not unique: the value $\theta + 2\pi n$, where $n = 0, \pm 1, \pm 2, \ldots$, can be taken instead of the value θ.

The trigonometric form is convenient for the operations of multiplication and division of complex numbers.

Suppose we are given two arbitrary complex numbers

$$z_1 = (x_1, y_1) = (\rho_1 \cos \theta_1, \rho_1 \sin \theta_1) \text{ and } z_2 = (x_2, y_2) = \\ = (\rho_2 \cos \theta_2, \rho_2 \sin \theta_2).$$

Then, by the definition of multiplication (in accordance with formula (7.2)), the product of these numbers has the form

$$z_1 \cdot z_2 = (x_1 x_2 - y_1 y_2, x_1 y_2 + x_2 y_1) = (\rho_1 \rho_2 \cos \theta_1 \cos \theta_2 - \\ - \rho_1 \rho_2 \sin \theta_1 \sin \theta_2, \rho_1 \rho_2 \cos \theta_1 \sin \theta_2 + \rho_1 \rho_2 \sin \theta_1 \cos \theta_2) = \\ = [(\rho_1 \rho_2) \cos (\theta_1 + \theta_2), (\rho_1 \rho_2) \sin (\theta_1 + \theta_2)]. \tag{7.7}$$

By analogy, we infer from formula (7.4) that the quotient z_1/z_2 of two complex numbers $z_1 = (x_1, y_1) = (\rho_1 \cos \theta_1, \rho_1 \sin \theta_1)$ and $z_2 = (x_2, y_2) = (\rho_2 \cos \theta_2, \rho_2 \sin \theta_2)$ has the form*

$$\frac{z_1}{z_2} = \left[\left(\frac{\rho_1}{\rho_2}\right) \cos (\theta_1 - \theta_2), \left(\frac{\rho_1}{\rho_2}\right) \sin (\theta_1 - \theta_2)\right]. \tag{7.8}$$

Formulas (7.7) and (7.8) lead to the conclusion that *when two complex numbers are multiplied their moduli are multiplied while the arguments are added together* (*when a complex number is divided by another complex number, the modulus of the dividend is divided by that of the divisor and the argument of the divisor is subtracted from the argument of the dividend*). This property holds true for the case of the product of any finite number of complex numbers. In particular, if n equal complex numbers are multiplied together (i.e. if a complex number is raised to the power n), then

$$(\rho \cos \theta, \rho \sin \theta)^n = (\rho^n \cos \theta n, \rho^n \sin \theta n). \tag{7.9}$$

From formula (7.9), for $\rho = 1$, we can obtain the so-called *De Moivre formula***

$$(\cos \theta, \sin \theta)^n = (\cos \theta n, \sin \theta n). \tag{7.10}$$

* In this case it is assumed that the complex number z_2 is nonzero, that is $\rho_2 \neq 0$.

** Abraham De Moivre (1667-1754), an English mathematician, a Frenchman by birth.

Formula (7.10) can be also represented as

$$(\cos\theta + i\sin\theta)^n = \cos\theta n + i\sin\theta n. \tag{7.11}$$

We should point out in conclusion that a complex number written in trigonometric form is equal to zero if and only if its modulus is equal to zero. This fact and the fact that when complex numbers are multiplied their moduli are multiplied together imply that *the product of several complex numbers is equal to zero only in the case when at least one of the factors is equal to zero.*

7.2. ALGEBRAIC POLYNOMIALS

7.2.1. An algebraic polynomial of the nth degree is an expression of the form

$$f(z) = c_0 z^n + c_1 z^{n-1} + \ldots + c_{n-1} z + c_n, \tag{7.12}$$

where $z = (x, y) = x + iy$ is a variable complex number, and $c_0, c_1, \ldots, c_n$ are some constant complex numbers the first of which is different from zero. As is known, any algebraic polynomial of the nth degree can be divided "in a column" by another algebraic polynomial of the degree not higher than n. Thus we arrive at the following assertion: *whatever the two polynomials $f(z)$ and $\varphi(z)$ such that the degree of $\varphi(z)$ is not higher than that of $f(z)$, there holds an equality*

$$f(z) = \varphi(z) \cdot q(z) + r(z), \tag{7.13}$$

where $q(z)$ and $r(z)$ are some polynomials, with the degree of $q(z)$ being equal to the difference between the degrees of the polynomials $f(z)$ and $\varphi(z)$ and the degree of $r(z)$ being lower than that of $\varphi(z)$.

The familiar terms "dividend", "divisor", "quotient" and "remainder" are usually used in connection with the polynomials $f(z)$, $\varphi(z)$, $q(z)$ and $r(z)$ entering into equality (7.13).

It is said that the polynomial $f(z)$ is divisible by the polynomial $\varphi(z)$ if in formula (7.13) resulting from a column division the remainder is $r(z) = 0$.

Let us agree to refer to any complex constant as a polynomial of degree zero. Then it is quite clear that *any polynomial is divisible by a nonzero polynomial of degree zero.* Let us consider the problem of divisibility of the polynomial $f(z)$ by a polynomial of the degree $(z - b)$.

Definition. *Let us call the complex number b a* **zero point** *(or a* **root***) of the polynomial $f(z)$ if $f(b)$ is equal to zero.*

Theorem 7.1. *The polynomial $f(z)$ of the zeroth degree is divisible by the binomial $(z - b)$ if and only if b is a root of the polynomial $f(z)$.*

Proof. Let us write formula (7.13) for the polynomials $f(z)$ and $\varphi(z) = (z - b)$. Since the degree of the remainder $r(z)$ in this formula *must be lower* than that of the divisor $\varphi(z) = z - b$, it follows that $r(z)$ is a *polynomial of degree zero*, i.e. $r(z) = c = \text{const}$. Thus, formula (7.13) assumes the form

$$f(z) = (z - b) \cdot q(z) + c. \tag{7.14}$$

Putting $z = b$ in formula (7.14), we shall find that $c = f(b)$. By definition $f(z)$ is divisible by $z - b$ if and only if the remainder $c = f(b)$ in formula (7.14) is equal to zero, that is, if and only if b is a zero point of $f(z)$. We have proved the theorem.

7.2.2. A natural question arises: does every algebraic polynomial possess zero points? The answer to this question is given by the *fundamental theorem of algebra**: *every polynomial of the nonzeroth degree has at least one zero point.*

Proceeding from this theorem we shall prove that an *algebraic polynomial of the nth degree has exactly n zero points***. Indeed, suppose $f(z)$ is a polynomial of the nth degree. In accordance with the fundamental theorem of algebra $f(z)$ possesses at least one root b_1, i.e. for $f(z)$ there holds a representation

$$f(z) = (z - b_1) f_1(z), \tag{7.15^1}$$

in which $f_1(z)$ denotes a certain polynomial of degree $(n - 1)$. If $n \neq 1$, then, according to the fundamental theorem of algebra, $f_1(z)$ has at least one root b_2, i.e. for $f_1(z)$ there holds a representation

$$f_1(z) = (z - b_2) f_2(z), \tag{7.15^2}$$

in which $f_2(z)$ denotes a certain polynomial of degree $(n - 2)$. Repeating the argument, we shall obtain representations

$$f_2(z) = (z - b_3) f_3(z), \tag{7.15^3}$$

$$\cdots\cdots\cdots$$

$$f_{n-1}(z) = (z - b_n) f_n(z). \tag{7.15^n}$$

In the last of these representations $f_n(z)$ denotes a certain polynomial of degree zero, i.e. $f_n(z) = c = \text{const}$. Comparing equalities (7.15^1)-(7.15^n) and taking into account that $f_n(z) = c$, we shall have

$$f(z) = (z - b_1)(z - b_2) \ldots (z - b_n) c. \tag{7.16}$$

---

* The proof of this theorem can be found in *The Theory of Functions of a Complex Variable* by A. Sveshnikov and A. Tikhonov, Mir Publishers, Moscow.

** In this case we assume, of course, that n is positive.

Note that the complex constant c is not equal to zero, for, if otherwise, the polynomial $f(z)$ would be identically equal to zero and would not be a polynomial of the nth degree*.

It is evident from equality (7.16) that $f(b_1) = f(b_2) = \ldots = f(b_n) = 0$, i.e. each of the numbers $b_1, b_2, \ldots, b_n$ is a root of the polynomial $f(z)$. Besides, it is evident from (7.16) that whatever the complex number b different from $b_1, b_2, \ldots, b_n$, the complex number $f(b)$ is not equal to zero**. Thus, the polynomial $f(z)$ possesses exactly n roots $b_1, b_2, \ldots, b_n$.

Equality (7.16) gives the factorization of the polynomial $f(z)$. If we know the form of the polynomial $f(z)$ (7.12), we can determine the constant c in equality (7.16). Comparing in equalities (7.16) and (7.12) the coefficients in z^n, we obtain $c = c_0$***.

Polynomial (7.12) in which $c_0 = 1$ is called a *reduced* polynomial. For a reduced polynomial the factorization formula (7.16) assumes the form

$$f(z) = (z - b_1)(z - b_2) \ldots (z - b_n). \tag{7.17}$$

Comparing formula (7.17) with formula (7.12) (for $c_0 = 1$), we obtain the following relations:

$$c_1 = -(b_1 + b_2 + \ldots + b_n),$$
$$c_2 = +(b_1b_2 + b_1b_3 + \ldots + b_{n-1}b_n)$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$c_n = (-1)^n b_1b_2 \ldots b_n.$$

In what follows, unless otherwise stated, we shall consider *reduced polynomials.*

7.3. MULTIPLE ROOTS OF POLYNOMIALS. TEST FOR DISTINGUISHING THE MULTIPLICITY OF A ROOT

There may happen to be *coinciding roots* among the roots of the polynomial $f(z)$. Let $a, b, \ldots, c$ be *distinct* roots of the reduced polynomial $f(z)$. Then, by virtue of the results obtained in the

* We use here the following assertion: *if the polynomial* $f(z) = a_0z^n + a_1z^{n-1} + \ldots + a_{n-1}z + a_n$ *is identically zero, then all its coefficients are equal to zero.* In fact, if $f(z) \equiv 0$, then for $z = 0$ we have $a_n = 0$. But then $f(z) \equiv z[a_0z^{n-1} + a_1z^{n-2} + \ldots + a_{n-1}] \equiv 0$. Since $z \not\equiv 0$, the expression in brackets is identically equal to zero, whence, for $z = 0$ we receive $a_{n-1} = 0$. Proceeding with an analogous argument, we shall prove that all the coefficients are equal to zero.

** Because the product of several complex numbers is equal to zero only if at least one of the factors is equal to zero (see 7.1).

*** Here we make use of the following assertion: *if two polynominals* $a_0z^n + a_1z^{n-1} + \ldots + a_n$ *and* $b_0z^n + b_1z^{n-1} + \ldots + b_n$ *are identically equal to each other, then* $a_0 = b_0, a_1 = b_1, \ldots, a_n = b_n$. To prove the assertion, it is sufficient to apply the assertion given in the first footnote on this page to the difference between the indicated polynomials.

previous section, for $f(z)$ there holds a factorization

$$f(z) = (z-a)^{\alpha}(z-b)^{\beta} \ldots (z-c)^{\gamma}. \tag{7.18}$$

In this factorization $\alpha, \beta, \ldots, \gamma$ are arbitrary integers, each of which is not smaller than unity, and $\alpha + \beta + \ldots + \gamma = n$, where n is the degree of the polynomial $f(z)$.

If factorization (7.18) *is valid for the polynomial* $f(z)$, *then the complex number* a *is said to be a zero of multiplicity* α, (*the* α*-fold root*) *of* $f(z)$, *the complex number* b *is said to be a zero of multiplicity* $\beta, \ldots$ *of* $f(z)$, *the complex number* c *is said to be a zero of multiplicity* γ *of* $f(z)$.

A root whose multiplicity is equal to inity is called a *simple* root, and that whose multiplicity exceeds unity is called a *multiple* root.

There can also be given another definition of a root of a given multiplicity which is equivalent to that cited above: *the complex number* a *is called a root of the polynomial* $f(z)$ *of multiplicity* α *if for* $f(z)$ *there holds a representation*

$$f(z) = (z-a)^{\alpha}\varphi(z), \text{ where } \varphi(a) \neq 0. \tag{7.19}$$

Our aim now is to indicate the necessary and sufficient condition for the complex number a to be a root of the polynomial $f(z)$ of multiplicity α.

Let us regard as the derivative of the polynomial $f(z)$ *the entire rational function* $f'(z)$ *obtained by means of a formal differentiation* * *of* $f(z)$ *with respect to* z. But first let us prove the following assertion.

Lemma 1. *If the complex number* a *is a root of multiplicity* α *of the polynomial* $f(z)$, *then this number* a *is at the same time a root of multiplicity* $(\alpha - 1)$ *of the polynomial* $f'(z)$.

Remark. In particular, for $\alpha = 1$ the number a, being a simple root of $f(z)$, is not a root of $f'(z)$.

Proof. By the hypothesis, representation (7.19) is valid for $f(z)$. Differentiating formula (7.19), we obtain

$$f'(z) = \alpha(z-a)^{\alpha-1}\varphi(z) + (z-a)^{\alpha}\varphi'(z),$$

or

$$f'(z) = (z-a)^{\alpha-1}\varphi_1(z), \tag{7.20}$$

where

$$\varphi_1(z) = \alpha\varphi(z) + (z-a)\varphi'(z).$$

Since $\varphi_1(a) = \alpha\varphi(a) \neq 0$, representation (7.20) means that the number a is a root of multiplicity $(\alpha - 1)$ of the polynomial $f'(z)$. We have proved the lemma.

---

* That is, the differentiation of $f(z)$ with respect to z is carried out as if z is a real variable.

Theorem 7.2. *For the complex number a to be a root of multiplicity* α *of the polynomial* $f(z)$, *it is necessary and sufficient that the following three conditions should be satisfied:*

$$f(a) = f'(a) = \ldots = f^{(\alpha-1)}(a) = 0, \quad f^{(\alpha)}(a) \neq 0. \tag{7.21}$$

Proof. (1) **Necessity.** Let a be a root of multiplicity α of the polynomial $f(z)$. Then, in accordance with Lemma 1, the same number a is a root of multiplicity $(\alpha - 1)$ of the polynomial $f'(z)$, a root of multiplicity $(\alpha - 2)$ of the polynomial $f^{(2)}(z)$, . . ., a root of multiplicity 1 of the polynomial $f^{(\alpha-1)}(z)$, i.e.

$$f(a) = f'(a) = \ldots = f^{(\alpha-1)}(a) = 0.$$

According to the remark to Lemma 1, the number a is not a root of the polynomial $f^{(\alpha)}(z)$, i.e. $f^{(\alpha)}(a) \neq 0$. We have proved that conditions (7.21) are satisfied.

(2) **Sufficiency.** Suppose conditions (7.21) are satisfied. It is required to prove that the number a is a root of multiplicity α of the polynomial $f(z)$. Since $f^{(\alpha-1)}(a) = 0$, the number a is a root of the polynomial $f^{(\alpha-1)}(z)$ *of multiplicity not lower than unity.* Thus, on the basis of Lemma 1, the number a is a root of the polynomial $f^{(\alpha-2)}(z)$ *of multiplicity not lower than* 2, a root of the polynomial $f^{(\alpha-3)}(z)$ *of multiplicity not lower than* 3, . . ., a root of the polynomial $f(z)$ *of multiplicity not lower than* α.

It remains to prove that the multiplicity of the root a of the polynomial $f(z)$ *does not exceed* α. Were the multiplicity higher than α, then, according to Lemma 1, the multiplicity of the root a of the polynomial $f^{(\alpha-1)}(z)$ would be higher than unity, whence it would follow that a is a root of $f^{(\alpha)}(z)$, i.e. $f^{(\alpha)}(a) = 0$, and this contradicts the last of the conditions (7.21). We have proved the theorem.

7.4. THE PRINCIPLE OF SEPARATION OF MULTIPLE ROOTS. EUCLIDEAN ALGORITHM

7.4.1. The principle of separation of multiple roots. Suppose our aim is to find for a given polynomial $f(z)$, having, in general, multiple roots, such a polynomial $F(z)$ which *has the same roots as* $f(z)$ *but all of them of multiplicity* 1. To achieve our aim, we shall introduce some new notions.

Definition 1. *We shall use the term* **common divisor** *of two polynomials* $f(z)$ *and* $\varphi(z)$ *for any polynomial by which both polynomials* $f(z)$ *and* $\varphi(z)$ *can be divided.*

Definition 2. *We shall use the term* **greatest common divisor** *of two polynomials* $f(z)$ *and* $\varphi(z)$ *for such their divisor which can be divided by any other divisor of the two polynomials.*

Let us agree to denote the greatest common divisor of the two polynomials $f(z)$ and $\varphi(z)$ by the symbol $D[f(z), \varphi(z)]$.

Note that the definition of the greatest common divisor implies that it is specified to within an arbitrary constant factor.

Returning to the aim formulated at the beginning of the present section, we can easily verify that the required polynomial $F(z)$ has the form

$$F(z) = \frac{f(z)}{D[f(z), f'(z)]}. \tag{7.22}$$

Indeed, assume

$$f(z) = (z-a)^{\alpha}(z-b)^{\beta} \ldots (z-c)^{\nu}, \tag{7.23}$$

where $a, b, \ldots, c$ are *distinct* roots. Then, according to Theorem 7.2, the following representation is valid for the polynomial $f'(z)$:

$$f'(z) = (z-a)^{\alpha-1}(z-b)^{\beta-1} \ldots (z-c)^{\nu-1}\psi(z), \tag{7.24}$$

where $\psi(z)$ does not contain the factors $(z-a), (z-b), \ldots, (z-c)$.

If we compare formulas (7.23) and (7.24), we shall see that

$$D[f(z), f'(z)] = (z-a)^{\alpha-1}(z-b)^{\beta-1} \ldots (z-c)^{\nu-1}. \tag{7.25}$$

In its turn, the comparison of formulas (7.23) and (7.25) shows that the polynomial $F(z)$ specified by formula (7.22) has the form

$$F(z) = (z-a)(z-b) \ldots (z-c). \tag{7.26}$$

Thus we have proved that the polynomial $F(z)$ specified by formula (7.22) posssesses the same roots as the polynomial $f(z)$, but all the roots are simple.

Thus, the problem of separating multiple roots reduces to constructing, from the given polynomial $f(z)$, the polynomial $F(z)$ specified by formula (7.22).

Since the denominator of equation (7.22) contains the greatest common divisor of two polynomials $f(z)$ and $f'(z)$ we are faced with the problem of finding the greatest common divisor of the two polynomials. Let us now try to solve this problem.

7.4.2. Finding the greatest common divisor of two polynomials (Euclidean algorithm). Suppose we are given two arbitrary polynomials $f(z)$ and $\varphi(z)$ and it is required to find their greatest common divisor. Without any loss of generality we shall assume that the degree of $\varphi(z)$ does not exceed that of $f(z)$. Then, having divided $f(z)$ by $\varphi(z)$ "in a column", we shall arrive at formula (7.13) (see 7.2)

$$f(z) = \varphi(z)\,q(z) + r_1(z), \tag{7.27^1}$$

in which, as was established in 7.2, the degree of the remainder $r_1(z)$ is lower than that of the divisor $\varphi(z)$. Consequently, we can

continue the process, dividing now $\varphi(z)$ by $r_1(z)$. As a result of this division we shall obtain a formula which is similar to formula (7.13):

$$\varphi(z) = r_1(z)\, q_1(z) + r_2(z), \tag{7.27^2}$$

where the degree of the remainder $r_2(z)$ is lower than that of the divisor $r_1(z)$.

We proceed with the operation, dividing, in a column, $r_1(z)$ by $r_2(z)$ and so on. The result is

$$r_1(z) = r_2(z)\, q_2(z) + r_3(z), \tag{7.27^3}$$

. . . . . .

$$r_{k-2}(z) = r_{k-1}(z)\, q_{k-1}(z) + r_k(z). \tag{7.27^k}$$

Since at each stage of the column division the degree of the remainder will be lower *at least by unity*, by repeating the process a sufficient number of times k, we shall obtain, at the $(k+1)$th stage, a remainder equal to zero*, i.e.

$$r_{k-1}(z) = r_k(z)\, q_k(z). \tag{7.27^{k+1}}$$

We shall now prove that the last nonzero remainder $r_k(z)$ is the *greatest common divisor* of the polynomials $f(z)$ and $\varphi(z)$.

It is sufficient to prove the following two assertions:

(1) the polynomials $f(z)$ and $\varphi(z)$ are divisible by $r_k(z)$ (this means that $r_k(z)$ is one of the divisors of $f(z)$ and $\varphi(z)$);

(2) the polynomial $r_k(z)$ is divisible by *any* divisor $r_0(z)$ of the polynomials $f(z)$ and $\varphi(z)$ (this means that $r_k(z)$ is the *greatest common divisor* of the indicated polynomials).

To prove assertion (1), we make note of the fact that, by virtue of (7.27^{k+1}), $r_{k-1}(z)$ is divisible by $r_k(z)$, and then, in accordance with (7.27^k), $r_{k-2}(z)$ is divisible by $r_k(z)$ . . . . Coming up the chain of equations (7.27^1)-(7.27^k), we shall finally prove that $\varphi(z)$ and $f(z)$ are divisible by $r_k(z)$.

Let us now prove assertion (2). Suppose $r_0(z)$ is any divisor of the polynomials $f(z)$ and $\varphi(z)$. By virtue of equality (7.27^1), $r_1(z)$ is divisible by $r_0(z)$, and then, in accordance with equation (7.27^2), $r_2(z)$ is divisible by $r_0(z)$, in accordance with (7.27^3) $r_3(z)$ is divisible by $r_0(z)$ . . . . Passing consecutively from equation (7.27^1) down to equation (7.27^k), we shall finally prove that $r_k(z)$ is divisible by $r_0(z)$

Thus we have completely substantiated the process, described

* If the remainder does not vanish at one of the intermediate stages of the process described, then, after some k stages, we shall obtain the remainder $r_k(z)$ of degree zero. Then, the next remainder $r_{k+1}(z)$ is obviously equal to zero (for any polynomial is divisible by a polynomial of degree zero).

above, of finding the greatest common divisor of two polynomials. This process is known as the *Euclidean algorithm.*

Example. Let us find the greatest common divisor of two polynomials*

$$f(z) = z^4 - 2z^3 + 3z^2 - 2z + 1 \text{ and } \varphi(z) = 4z^3 - 6z^2 + 6z - 2.$$

Dividing $f(z)$ by $\varphi(z)$ in a column, we obtain

$$\begin{array}{l|l} z^4 - 2z^3 + 3z^2 - 2z + 1 & 4z^3 - 6z^2 + 6z - 2 \\ \underline{z^4 - \frac{3}{2}z^3 + \frac{3}{2}z^2 - \frac{1}{2}z} & \underline{\frac{1}{4}z - \frac{1}{8}} \\ -\frac{1}{2}z^3 + \frac{3}{2}z^2 - \frac{3}{2}z + 1 & \\ \underline{-\frac{1}{2}z^3 + \frac{3}{2}z^2 - \frac{3}{2}z + \frac{1}{4}} & \\ \boxed{\frac{3}{4}z^2 - \frac{3}{4}z + \frac{3}{4}} & \end{array}$$

We should have then divided $\varphi(z)$ by the polynomial given in a dotted frame. However, since the greatest common divisor has been *determined with an accuracy to within an arbitrary constant factor*, it is more convenient to multiply the remainder in a dotted frame by $\frac{4}{3}$ and then divide $\varphi(z)$ by the polynomial $z^2 - z + 1$. The result will be

$$\begin{array}{l|l} 4z^3 - 6z^2 + 6z - 2 & z^2 - z + 1 \\ \underline{4z^3 - 4z^2 + 4z} & \underline{4z - 2} \\ -2z^2 + 2z - 2 & \\ \underline{-2z^2 + 2z - 2} & \\ \text{the remainder is zero.} & \end{array}$$

Thus, the greatest common divisor of the polynomials $f(z)$ and $\varphi(z)$ is equal to $z^2 - z + 1$, i.e.

$$D[f(z), \varphi(z)] = z^2 - z + 1.$$

Remark 1. For the sake of simplicity, the polynomials $f(z)$ and $\varphi(z)$ in the above example are taken with *real* coefficients. The same methods hold true for polynomials with complex coefficients.

Remark 2. It should be pointed out that up till now there have not been found stable numerical methods of estimating the roots of arbitrary polynomials with the required degree of accuracy. However, knowing the position of the sought-for root on some seg-

* It is easy to see that $\varphi(z) = f'(z)$.

ment of the number axis, we can calculate it with the required accuracy by means of the methods described in 12.1.

7.5. DECOMPOSITION OF A PROPER RATIONAL FRACTION WITH COMPLEX COEFFICIENTS INTO A SUM OF PARTIAL FRACTIONS

A **rational fraction** is a ratio of two algebraic polynomials. A rational fraction is called *proper* if the degree of the polynomial in the numerator is lower than the degree of the polynomial in the denominator. Otherwise, the rational fraction is called *improper*. As a rule, we shall designate a rational fraction as $P(z)/Q(z)$, assuming $P(z)$ and $Q(z)$ to be algebraic polynomials.

Lemma 2. *Assume $P(z)/Q(z)$ to be a proper rational fraction whose denominator has a complex number a as its root of multiplicity α, i.e.*

$$Q(z) = (z-a)^{\alpha}\varphi(z), \text{ where } \varphi(a) \neq 0. \tag{7.28}$$

Then, the following representation is valid for this fraction:

$$\frac{P(z)}{Q(z)} = \frac{A}{(z-a)^{\alpha}} + \frac{\psi(z)}{(z-a)^{\alpha-k}\varphi(z)}. \tag{7.29}$$

In this representation A is a complex constant equal to $A = \frac{P(a)}{\varphi(a)}$, k is an integer equal to or greater than unity, and $\psi(z)$ is some polynomial, the last fraction on the right-hand side of (7.29) being proper.

Proof. Denoting by A a constant number of the kind $A = \frac{P(a)^*}{\varphi(a)}$, let us consider the difference

$$\frac{P(z)}{Q(z)} - \frac{A}{(z-a)^{\alpha}}.$$

Reducing this difference to a common denominator, we obtain

$$\frac{P(z)}{Q(z)} - \frac{A}{(z-a)^{\alpha}} = \frac{P(z)-A\varphi(z)}{(z-a)^{\alpha}\varphi(z)} = \frac{\Phi(z)}{(z-a)^{\alpha}\varphi(z)}, \tag{7.30}$$

where $\Phi(z)$ denotes a polynomial of the kind $\Phi(z) = P(z) - A\varphi(z)$. Because of the equality to zero of the polynomial $\Phi(a) = P(a) - A\varphi(a)$, the complex number *$a$ is a zero of the polynomial* $\Phi(z)$ of some multiplicity $k \geqslant 1$, i.e.

$$\Phi(z) = (z-a)^k\psi(z), \text{ where } \psi(a) \neq 0. \tag{7.31}$$

Substituting representation (7.31) into formula (7.30), we receive

$$\frac{P(z)}{Q(z)} - \frac{A}{(z-a)^{\alpha}} = \frac{\psi(z)}{(z-a)^{\alpha-k}\varphi(z)}. \tag{7.32}$$

* The number A is meaningful since $\varphi(a) \neq 0$ by virtue of (7.28).

Thus we have proved formula (7.29). It remains to make sure that the fraction on the right-hand side of (7.32) is proper. The validity of this statement follows directly from the fact that *the difference between two proper fractions is a proper fraction**.

We have proved Lemma 2.

Lemma 2 implies the following remarkable theorem establishing the fact that a proper rational fraction can be decomposed into a sum of partial fractions.

Theorem 7.3. *Assume $P(z)/Q(z)$ to be a proper rational fraction, whose denominator has the form*

$$Q(z) = (z-a)^{\alpha}(z-b)^{\beta}\ldots(z-c)^{\gamma}. \tag{7.33}$$

Then, the following representation is valid for this fraction:

$$\begin{aligned}\frac{P(z)}{Q(z)} = {} & \frac{A_1}{(z-a)^{\alpha}} + \frac{A_2}{(z-a)^{\alpha-1}} + \ldots + \frac{A_{\alpha}}{(z-a)} + \\ & + \frac{B_1}{(z-b)^{\beta}} + \frac{B_2}{(z-b)^{\beta-1}} + \ldots + \frac{B_{\beta}}{(z-b)} + \\ & + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + \\ & + \frac{C_1}{(z-c)^{\gamma}} + \frac{C_2}{(z-c)^{\gamma-1}} + \ldots + \frac{C_{\gamma}}{(z-c)}.\end{aligned} \tag{7.34}$$

In this representation $A_1, A_2, \ldots, A_{\alpha}, B_1, B_2, \ldots, B_{\beta}, \ldots \ldots, C_1, C_2, \ldots, C_{\gamma}$ are arbitrary constant complex numbers, some of which can be equal to zero.

Proof. We shall first apply Lemma 2 to the fraction $P(z)/Q(z)$ bearing in mind that the complex number a is a root of $Q(z)$ of multiplicity α. We shall obtain equality (7.32). To the right-hand side of this equality we shall again apply Lemma 2, keeping in mind that either the complex number a is an $(\alpha - k)$-fold root of the denominator on the indicated right-hand side (for $\alpha - k > 0$), or, by virtue of decomposition (7.33), the complex number b is a β-fold root of this denominator (for $\alpha - k \leqslant 0$). As a result, we shall obtain an equality, analogous to (7.32), to the right-hand side of which we can again apply Lemma 2. Proceeding with this argument (that is, applying consecutively Lemma 2 to all the roots of $Q(z)$), we shall have representation (7.34) for the fraction $P(z)/Q(z)$. The proof of the theorem is complete.

Remark. Since the number k in Lemma 2 may be larger than unity and the polynomial $P(z)$ may have roots coinciding with those of $Q(z)$, a part of the coefficients $A_1, \ldots, A_{\alpha}, B_1, \ldots, B_{\beta}, C_1, \ldots \ldots, C_{\gamma}$ in formula (7.34) may be equal to zero.

---

* We can easily ascertain this by reducing the difference between proper fractions to a common denominator.

7.6. FACTORIZATION OF AN ALGEBRAIC POLYNOMIAL WITH REAL COEFFICIENTS INTO A PRODUCT OF IRREDUCIBLE REAL FACTORS

In the previous section we studied the decomposition of a rational fraction with *complex* coefficients into a sum of partial fractions. Our end purpose is to decompose a rational fraction with *real* coefficients into a sum of partial fractions with *real* coefficients.

To achieve this purpose we must, first of all, determine the factorization of an algebraic polynomial with real coefficients into a product of irreducible real factors. This is precisely the subject matter of the present section.

Assume that

$$f(z) = z^n + c_1 z^{n-1} + c_2 z^{n-2} + \ldots + c_n \tag{7.35}$$

is a reduced algebraic polynomial with real coefficients $c_1, c_2, \ldots, c_n$.

Let us first prove the following theorem.

Theorem 7.4. *If a complex number a is a root of an algebraic polynomial with real coefficients* (7.35), *then its conjugate complex number** $\bar{a}$ *is also a root of the polynomial* (7.35). *Moreover, if the complex root a is of multiplicity λ, then the root $\bar{a}$ is of multiplicity λ too.*

Proof. First of all let us prove the following auxiliary assertion: if $f(z)$ is a polynomial with real coefficients, then the complex quantity $f(\bar{z})$ is the conjugate of $f(z)$. It is sufficient to prove that for any number n the quantity $(\bar{z})^n$ is the conjugate of the quantity z^n. This latter fact immediately follows from the trigonometric form of a complex number. Indeed, assume

$$z = \rho(\cos\theta + i\sin\theta).$$

Then

$$\bar{z} = \rho[\cos(-\theta) + i\sin(-\theta)].$$

By virtue of the De Moivre formula (7.11)

$$z^n = \rho^n(\cos\theta n + i\sin\theta n),$$

$$(\bar{z})^n = \rho^n[\cos(-\theta n) + i\sin(-\theta n)] = \rho^n(\cos\theta n - i\sin\theta n).$$

The comparison of the last two formulas shows that $(\bar{z})^n$ is the conjugate of z^n. We have thus proved the auxiliary assertion.

Suppose now the complex number a is a root of the polynomial $f(z)$, i.e. $f(a) = 0$. We have established in 7.1 that a complex

* In what follows we shall designate the complex number conjugate of the given number by the same symbol as the given number but with a bar over it.

number is equal to zero if and only if its conjugate is equal to zero. Consequently, the equality $f(a) = 0$ and the auxiliary assertion proved above imply that $f(\bar{a}) = 0$, that is, the number $\bar{a}$ is a root of $f(z)$.

Assume that the root a is of multiplicity λ. Then, in accordance with Theorem 7.2

$$f(a) = f'(a) = f^{(2)}(a) = \ldots = f^{(\lambda-1)}(a) = 0;\ f^{(\lambda)}(a) \neq 0. \tag{7.36}$$

Since a complex number is equal to zero if and only if its conjugate is equal to zero, the auxiliary assertion proved above and relation (7.26) yield the following relations*:

$$f(\bar{a}) = f'(\bar{a}) = f^{(2)}(\bar{a}) = \ldots = f^{(\lambda-1)}(\bar{a}) = 0,\quad f^{(\lambda)}(\bar{a}) \neq 0. \tag{7.37}$$

By virtue of Theorem 7.2, relations (7.37) mean that the number $\bar{a}$ is a λ-fold root of $f(z)$. We have thus proved Theorem 7.4.

With the aid of Theorem 7.4 we shall determine the factorization of the polynomial $f(x)$ with real coefficients** into a product of irreducible real factors. Assume that the polynomial $f(x)$ has real roots $b_1, b_2, \ldots, b_m$ of multiplicity $\beta_1, \beta_2, \ldots, \beta_m$ respectively, and complex-conjugate pairs of roots a_1 and $\bar{a}_1$, a_2 and $\bar{a}_2, \ldots, a_n$ and $\bar{a}_n$, each pair of multiplicity $\lambda_1, \lambda_2, \ldots, \lambda_n$ respectively.

Then, in accordance with the results obtained in 7.3, the polynomial $f(x)$ can be represented as

$$f(x) = (x-b_1)^{\beta_1}(x-b_2)^{\beta_2}\ldots(x-b_m)^{\beta_m}(x-a_1)^{\lambda_1}\times$$
$$\times(x-\bar{a}_1)^{\lambda_1}(x-a_2)^{\lambda_2}(x-\bar{a}_2)^{\lambda_2}\ldots(x-a_n)^{\lambda_n}(x-\bar{a}_n)^{\lambda_n}. \tag{7.38}$$

Let us denote the real part and the imaginary part of the root a_k ($k = 1, 2, \ldots, n$) by u_k and v_k respectively, i.e. assume that $a_k = u_k + iv_k$. Then $\bar{a}_k = u_k - iv_k$. Let us transform, for any $k = 1, 2, \ldots, n$, the expression

$$(x-a_k)^{\lambda_k}(x-\bar{a}_k)^{\lambda_k} = [(x-a_k)(x-\bar{a}_k)]^{\lambda_k} =$$
$$= [(x-u_k-iv_k)(x-u_k+iv_k)]^{\lambda_k} = [(x-u_k)^2+v_k^2]^{\lambda_k} =$$
$$= (x^2+p_kx+q_k)^{\lambda_k}, \tag{7.39}$$

where

$$p_k = -2u_k,\quad q_k = u_k^2+v_k^2.$$

---

* We take into account that the derivative of a polynomial with real coefficients is also a polynomial with real coefficients.

** In what follows we shall deal with polynomials in a variable taking on *only real values*. Therefore, it is more convenient to denote the variable by the letter x and not by z.

Substituting (7.39) into (7.38), we shall finally obtain the following factorization of the polynomial $f(x)$ into a product of real irreducible factors:

$$f(x) = (x-b_1)^{\beta_1}(x-b_2)^{\beta_2}\ldots(x-b_m)^{\beta_m}(x^2+p_1x+q_1)^{\lambda_1}\times$$
$$\times(x^2+p_2x+q_2)^{\lambda_2}\ldots(x^2+p_nx+q_n)^{\lambda_n}. \qquad (7.40)$$

We infer that the polynomial $f(x)$ with real coefficients can be factorized into the product (7.40) of irreducible real factors, with the factors corresponding to the real roots having the form of binomials in degrees equal to the multiplicity of the roots and the factors corresponding to the complex pairs of roots having the form of quadratic trinomials in degrees equal to the multiplicity of these pairs of roots.

7.7. DECOMPOSITION OF A PROPER RATIONAL FRACTION WITH REAL COEFFICIENTS INTO A SUM OF PARTIAL FRACTIONS WITH REAL COEFFICIENTS

The following two assertions hold true:

Lemma 3. *Assume that $P(x)/Q(x)$ is a proper rational fraction with real coefficients, whose denominator has a real number a as its root of multiplicity α, i.e.*

$$Q(x) = (x-a)^{\alpha}\varphi(x), \text{ where } \varphi(a) \neq 0.$$

Then, the following representation is valid for this fraction:

$$\frac{P(x)}{Q(x)} = \frac{A}{(x-a)^{\alpha}} + \frac{\psi(x)}{(x-a)^{\alpha-k}\varphi(x)}. \qquad (7.41)$$

In this representation A is a real number equal to $P(a)/\varphi(a)$, k is an integer equal to or greater than unity, and $\psi(x)$ is some polynomial with real coefficients, with the last fraction on the right-hand side of (7.41) *being proper.* It is not necessary to prove Lemma 3 since it follows directly from Lemma 2. We should only pay attention to the fact that since $P(x)$ and $Q(x)$ are polynomials with real coefficients and a is a real root, the polynomials $\varphi(x)$ and $\psi(x)$ also have real coefficients and, consequently, the constant number $A = P(a)/\varphi(a)$ is real.

Lemma 4. *Assume that $\frac{P(x)}{Q(x)}$ is a proper rational fraction with real coefficients whose denominator $Q(x)$ has complex numbers $a = u + iv$ and $\bar{a} = u - iv$ as its roots of multiplicity λ, i.e.*

$$Q(x) = (x^2+px+q)^{\lambda}\varphi(x), \text{ where } \varphi(a) \neq 0,$$
$$\varphi(\bar{a}) \neq 0,\ p = -2u,\ q = u^2+v^2. \qquad (7.42)$$

Then, the following representation is valid for this fraction:

$$\frac{P(x)}{Q(x)} = \frac{Mx+N}{(x^2+px+q)^\lambda} + \frac{\psi(x)}{(x^2+px+q)^{\lambda-k}(x)}. \tag{7.43}$$

In this representation M and N are some real constants, k is an integer $\geqslant 1$, and $\psi(x)$ is some polynomial with real coefficients, the last fraction on the right-hand side of (7.43) being proper.

Proof of Lemma. 4. Let us agree to denote the real part of the complex quantity A by the symbol Re $[A]$ and the imaginary part of the complex quantity A by the symbol Im $[A]$. Let us put*

$$M = \frac{1}{v}\operatorname{Im}\left[\frac{P(a)}{\varphi(a)}\right], \quad N = \operatorname{Re}\left[\frac{P(a)}{\varphi(a)}\right] - \frac{u}{v}\operatorname{Im}\left[\frac{P(a)}{\varphi(a)}\right].$$

It is easy to verify that the indicated M and N are the solution of the following equation:

$$P(a) - (Ma + N)\varphi(a) = 0. \tag{7.44}$$

Indeed, having divided this equation by $\varphi(a)$ and having equated the real and imaginary parts to zero, we obtain two equations

$$\left.\begin{aligned} Mu + N &= \operatorname{Re}\left[\frac{P(a)}{\varphi(a)}\right], \\ Mv &= \operatorname{Im}\left[\frac{P(a)}{\varphi(a)}\right], \end{aligned}\right\}$$

from which we can determine M and N written out above. Let us now consider the difference

$$\frac{P(x)}{Q(x)} - \frac{Mx+N}{(x^2+px+q)^\lambda}.$$

Reducing this difference to a common denominator, we receive

$$\frac{P(x)}{Q(x)} - \frac{Mx+N}{(x^2+px+q)^\lambda} = \frac{P(x)-(Mx+N)\varphi(x)}{(x^2+px+q)^\lambda\varphi(x)} = \frac{\Phi(x)}{(x^2+px+q)^\lambda\varphi(x)}. \tag{7.45}$$

Here $\Phi(x)$ denotes a polynomial with real coefficients of the form $\Phi(x) = P(x) - (Mx+N)\varphi(x)$. Equality (7.44) admits of the assertion that the complex number a, and, consequently, by virtue of Theorem 7.4, its conjugate number $\bar{a}$, are roots of the polynomial $\Phi(x)$ of some multiplicity $k \geqslant 1$. In that case, for the polynomial $\Phi(x)$, there holds a representation

$$\Phi(x) = (x^2 + px + q)^k\psi(x), \tag{7.46}$$

* Since $\varphi(a) \neq 0$ because of (7.42), we can rightfully consider the ratio $\frac{P(a)}{\varphi(a)}$.

where $\psi(x)$ is some polynomial with real coefficients, which does not have the numbers a and $\bar{a}$ as its roots. Substituting representation (7.46) into formula (7.45), we obtain representation (7.43). The fact that the last fraction on the right-hand side of (7.43) is proper follows from the equality of this fraction to the difference between two proper fractions.

We have proved Lemma 4.

A successive application of Lemmas 3 and 4 to the fraction $\frac{P(x)}{Q(x)}$ with respect to all the roots of the denominator leads to the following remarkable assertion.

Theorem 7.5. *Assume that $\frac{P(x)}{Q(x)}$ is a proper rational fraction with real coefficients, whose denominator has the form*

$$Q(x) = (x-b_1)^{\beta_1}(x-b_2)^{\beta_2}\ldots(x-b_m)^{\beta_m}(x^2+p_1x+q_1)^{\lambda_1}\ldots$$
$$\ldots(x^2+p_nx+q_n)^{\lambda_n}.$$

Then the following decomposition of this fraction into a sum of partial fractions is valid:

$$\frac{P(x)}{Q(x)} = \frac{B_1^{(1)}}{(x-b_1)} + \frac{B_2^{(1)}}{(x-b_1)^2} + \ldots + \frac{B_{\beta_1}^{(1)}}{(x-b_1)^{\beta_1}} + \ldots + \frac{B_1^{(m)}}{(x-b_m)} +$$
$$+ \frac{B_2^{(m)}}{(x-b_m)^2} + \ldots + \frac{B_{\beta_m}^{(m)}}{(x-b_m)^{\beta_m}} + \frac{M_1^{(1)}x+N_1^{(1)}}{(x^2+p_1x+q_1)} +$$
$$+ \frac{M_2^{(1)}x+N_2^{(1)}}{(x^2+p_1x+q_1)^2} + \ldots + \frac{M_{\lambda_1}^{(1)}+N_{\lambda_1}^{(1)}}{(x^2+p_1x+q_1)^{\lambda_1}} + \ldots$$
$$\ldots + \frac{M_1^{(n)}x+N_1^{(n)}}{(x^2+p_nx+q_n)} + \frac{M_2^{(n)}x+N_2^{(n)}}{(x^2+p_nx+q_n)^2} + \ldots + \frac{M_{\lambda_n}^{(n)}x+N_{\lambda_n}^{(n)}}{(x^2+p_nx+q_n)^{\lambda_n}}. \tag{7.47}$$

In this decomposition $B_1^{(1)}, B_2^{(1)}, \ldots, B_{\beta_m}^{(m)}, M_1^{(1)}, N_1^{(1)}, \ldots, M_{\lambda_n}^{(n)}, N_{\lambda_n}^{(n)}$ are arbitrary real constants some of which can be equal to zero.

Remark. For a concrete specification of the constant indicated above, equality (7.47) should be reduced to a common denominator and then the coefficients in the same degree of x should be compared in the numerators.

Examples and explanations.

1°. Decompose the proper fraction

$$\frac{2x^3+4x^2+x+2}{(x-1)^2(x^2+x+1)}$$

into a sum of partial fractions. Having ascertained that the quadratic trinomial x^2+x+1 has complex roots, we seek, in accordance

with Theorem 7.5, a decomposition of the form

$$\frac{2x^3+4x^2+x+2}{(x-1)^2(x^2+x+1)}=\frac{B_1}{x-1}+\frac{B_2}{(x-1)^2}+\frac{Mx+N}{x^2+x+1}. \tag{7.48}$$

Reducing equation (7.48) to a common denominator, we obtain

$$\frac{2x^3+4x^2+x+2}{(x-1)^2(x^2+x+1)}=\frac{B_1(x^3-1)+B_2(x^2+x+1)+(Mx+N)(x^2-2x+1)}{(x-1)^2(x^2+x+1)}.$$

Comparing in the numerators the coefficients in x^0, x^1, x^2 and x^3, we arrive at a system of equations*

$$\left.\begin{aligned} &B_1+M=2,\\ &B_2+N-2M=4,\\ &B_2+M-2N=1,\\ &-B_1+B_2+N=2. \end{aligned}\right\}$$

Having solved the system, we find that $B_1 = 2$, $B_2 = 3$, $M = 0$, $N = 1$. We finally obtain

$$\frac{2x^3+4x^2+x+2}{(x-1)^2(x^2+x+1)}=\frac{2}{x-1}+\frac{3}{(x-1)^2}+\frac{1}{x^2+x+1}. \tag{7.49}$$

The method of seeking the decomposition of a proper rational fraction we have just illustrated is known as the *method of undetermined coefficients*. This method is always justified; it is not necessary to prove the solvability of the system of equations resulting from the application of this method, the solvability follows from Theorem 7.5.

2°. We shall illustrate the method of undetermined coefficients by another example. It is required to find the decomposition of the proper fraction

$$\frac{3x^4+2x^3+3x^2-1}{(x-2)(x^2+1)^2}.$$

Since the quadratic trinomial $x^2 + 1$ has complex roots, we seek, in accordance with Theorem 7.5, a decomposition of the form

$$\frac{3x^4+2x^3+3x^2-1}{(x-2)(x^2+1)^2}=\frac{B}{x-2}+\frac{M_1x+N_1}{x^2+1}+\frac{M_2x+N_2}{(x^2+1)^2}.$$

We reduce the last equation to a common denominator and then compare the numerators. We receive

$$3x^4+2x^3+3x^2-1=B(x^4+2x^2+1)+$$
$$+(M_1x+N_1)(x^3-2x^2+x-2)+(M_2x+N_2)(x-2).$$

* In this case we use the assertion formulated in the footnote on p. 203.

Comparing the coefficients in x^0, x^1, x^2, x^3 and x^4, we arrive at a system of equations

$$\left.\begin{aligned} &B+M_1=3,\\ &N_1-2M_1=2,\\ &2B+M_1-2N_1+M_2=3,\\ &N_1-2M_1+N_2-2M_2=0,\\ &B-2N_1-2N_2=-1. \end{aligned}\right\}$$

Solving this system, we obtain $B = 3$, $M_1 = 0$, $N_1 = 2$, $M_2 = 1$, $N_2 = 0$. The final result is

$$\frac{3x^4+2x^3+3x^2-1}{(x-2)(x^2+1)^2}=\frac{3}{x-2}+\frac{2}{x^2+1}+\frac{x}{(x^2+1)^2}. \tag{7.50}$$

3°. As can be seen from the examples considered, the method of undetermined coefficients is rather cumbersome. It is natural, therefore, in the cases where it is possible, to seek some other, more simple, method of finding coefficients in the decomposition of a proper rational fraction into a sum of partial fractions. Suppose the denominator $Q(x)$ of the proper rational fraction $P(x)/Q(x)$ has the real number a as its α-fold root. Then among the partial fractions into whose sum the fraction $P(x)/Q(x)$ has been decomposed will be a fraction

$$\frac{A}{(x-a)^\alpha}. \tag{7.51}$$

We shall indicate a very simple method of computing the coefficient A in this partial fraction. Making use of Lemma 3 and formula (7.41), we shall ascertain that the coefficient A is equal to

$$A=\frac{P(a)}{\varphi(a)}, \quad \text{where} \quad \varphi(x)=\frac{Q(x)}{(x-a)^\alpha}.$$

We arrive at the following rule: *to compute the coefficient A in the partial fraction* (7.51), *corresponding to the real root a of the polynomial $Q(x)$ of multiplicity α, we must delete the expression* $(x-a)^\alpha$ *in the denominator of the fraction* $\frac{P(x)}{Q(x)}$ *and put* $x=a$ *in the remaining expression.*

The indicated method of finding the coefficient A is usually called the *deletion method.* Note that this method can be only employed to compute the coefficients in the *higher degrees of partial fractions, corresponding to the real roots of* $Q(x)$.

The method of deletion is especially efficient in the case where the denominator $Q(x)$ *has only simple real roots*, i.e. where $Q(x) = (x-a_1)(x-a_2)\ldots(x-a_n)$. Then, as we know, there holds

a decomposition

$$\frac{P(x)}{Q(x)}=\frac{A_1}{x-a_1}+\frac{A_2}{x-a_2}+\ldots+\frac{A_k}{x-a_k}+\ldots+\frac{A_n}{x-a_n},$$

in which all the coefficients can be computed with the aid of the deletion method. To compute the coefficient A_k, we must delete the parenthesis $(x-a_k)$ in the denominator of the fraction $P(x)/Q(x)$ and put $x=a_k$ in the remaining expression.

Example. Find the decomposition of the fraction

$$\frac{x+1}{(x-1)\,x\,(x-2)}. \tag{7.52}$$

In accordance with Theorem 7.5 we write

$$\frac{x+1}{(x-1)\,x\,(x-2)}=\frac{A_1}{x-1}+\frac{A_2}{x}+\frac{A_3}{x-2}.$$

To determine A_1, we delete the parenthesis $(x-1)$ in (7.52) and take $x=1$ in the remaining expression. We obtain $A_1=-2$. In a similar way we can find $A_2=1/2$, $A_3=3/2$.

The final result is

$$\frac{x+1}{(x-1)\,x\,(x-2)}=\frac{-2}{x-1}+\frac{1}{2x}+\frac{3}{2(x-2)}. \tag{7.53}$$

7.8. THE PROBLEM OF INTEGRATING RATIONAL FRACTIONS

At this stage we are already able to solve, in the general form, the problem of integrating rational fractions with real coefficients.

It should be emphasized, first of all, that this problem reduces to the problem of integrating *only a proper* rational fraction, for every improper rational fraction can be represented (by means of a column division of the numerator by the denominator) as the sum of an algebraic polynomial and a proper rational fraction.

Example.

$$\frac{x^4-x^3+1}{x^2+x+2}=(x^2-2x)+\frac{4x+1}{x^2+x+2},$$

since

$$\begin{array}{l|l} x^4-x^3+1 & x^2+x+2 \\ \underline{x^4+x^3+2x^2} & \overline{x^2-2x}. \\ -2x^3-2x^2+1 & \\ \underline{-2x^3-2x^2-4x} & \\ \text{remainder } 1+4x & \end{array}$$

We know how to integrate a polynomial (recall that an indefinite integral of a polynomial is a polynomial of a degree higher by unity).

It remains to learn to integrate a *proper* rational fraction. By virtue of Theorem 7.5, the problem of integrating a proper rational fraction reduces to integrating partial fractions of the *following four types*:

$$\text{I. } \frac{B}{x-b}, \quad \text{II. } \frac{B}{(x-b)^{\beta}}, \quad \text{III. } \frac{Mx+N}{(x^2+px+q)}, \quad \text{IV. } \frac{Mx+N}{(x^2+px+q)^{\lambda}}. \tag{7.54}$$

Here $\beta = 2, 3, \ldots$; $\lambda = 2, 3, \ldots$; B, M, N, b, p, and q are some real numbers, the trinomial x^2+px+q having no real roots, i.e. $q-\frac{p^2}{4}>0$.

Let us prove that each of the indicated four fractions is integrable in terms of elementary functions.

Fractions of types I and II can be integrated in an elementary way by means of the substitution $t = x - b$. We obtain

$$\int \frac{B}{x-b}\,dx = B\int \frac{dt}{t} = B\ln|t|+C = B\ln|x-b|+C, \tag{7.55}$$

$$\int \frac{B}{(x-b)^{\beta}}\,dx = B\int \frac{dt}{t^{\beta}} = -\frac{B}{(\beta-1)}\frac{1}{t^{\beta-1}}+C = \frac{-B}{(\beta-1)}\frac{1}{(x-b)^{\beta-1}}+C. \tag{7.56}$$

To integrate a fraction of type III, we shall represent the quadratic trinomial in the form $(x^2+px+q)=\left(x+\frac{p}{2}\right)^2+\left(q-\frac{p^2}{4}\right)$ and, taking into account that $q-\frac{p^2}{4}>0$, shall introduce into consideration a real constant $a=+\sqrt{q-\frac{p^2}{4}}$. Having made the substitution $t=x+\frac{p}{2}$, we obtain

$$\int \frac{(Mx+N)}{(x^2+px+q)}\,dx = \int \frac{Mt+\left(N-\frac{Mp}{2}\right)}{t^2+a^2}\,dt =$$

$$= \frac{M}{2}\int \frac{2t\,dt}{t^2+a^2}+\left(N-\frac{Mp}{2}\right)\int \frac{dt}{t^2+a^2} =$$

$$= \frac{M}{2}\int \frac{d(t^2+a^2)}{t^2+a^2}+\left(N-\frac{Mp}{2}\right)\frac{1}{a}\int \frac{d\left(\frac{t}{a}\right)}{\left(\frac{t}{a}\right)^2+1} =$$

$$= \frac{M}{2}\ln(t^2+a^2)+\frac{2N-Mp}{2a}\arctan\frac{t}{a}+C =$$

$$= \frac{M}{2}\ln(x^2+px+q)+\frac{2N-Mp}{2\sqrt{q-\frac{p^2}{4}}}\arctan\frac{x+\frac{p}{2}}{\sqrt{q-\frac{p^2}{4}}}+C. \tag{7.57}$$

It remains to integrate a fraction of type IV. Making use of the designations $t=x+\frac{p}{2}$, $a=\sqrt{q-\frac{p^2}{4}}$ introduced above, we receive

$$\int \frac{Mx+N}{(x^2+px+q)^\lambda}\,dx=\int \frac{Mt+\left(N-\frac{Mp}{2}\right)}{(t^2+a^2)^\lambda}\,dt=$$
$$=\frac{M}{2}\int \frac{d(t^2+a^2)}{(t^2+a^2)^\lambda}+\left(N-\frac{Mp}{2}\right)\int \frac{dt}{(t^2+a^2)^\lambda}.$$

Let us introduce the notation

$$I=\int \frac{d(t^2+a^2)}{(t^2+a^2)^\lambda},$$
$$K_\lambda=\int \frac{dt}{(t^2+a^2)^\lambda}.$$

The integral we are interested in will result from the calculations of the integrals I and K_λ. The computation of integral I is elementary:

$$I=-\frac{1}{(\lambda-1)}\frac{1}{(t^2+a^2)^{\lambda-1}}+C=-\frac{1}{(\lambda-1)}\frac{1}{(x^2+px+q)^{\lambda-1}}+C.$$

Integral K_λ was calculated in Example 6 at the end of 6.2. We obtained there, for this integral, recurrence formula (6.12), making it possible to calculate K_λ successively for any $\lambda = 2, 3, \ldots$ proceeding from the fact that

$$K_1=\int \frac{dt}{t^2+a^2}=\frac{1}{a}\arctan\frac{t}{a}+C.$$

Thus, we have found integrals of the four partial fractions (7.54) and proved that each of these integrals is an *elementary function**. And so we have arrived at the following theorem exhausting the problem of integrating a rational fraction.

Theorem 7.6. *Every rational fraction is integrable in terms of elementary functions.*

To conclude this section, we shall consider some examples of computing indefinite integrals of rational fractions. Let us calculate the indefinite integrals of the three fractions discussed in the previous section, (7.49), (7.50) and (7.53). Using these three formulas as

* To be more precise, it is expressed through a logarithm, an arc tangent and a rational function.

well as formulas (7.55), (7.56) and (7.57), we obtain

1. $\int \frac{2x^3+4x^2+x+2}{(x-1)^2(x^2+x+1)}dx = \int \frac{2}{x-1}dx + \int \frac{3}{(x-1)^2}dx + \int \frac{dx}{x^2+x+1} =$
$= 2\ln|x-1| - \frac{3}{x-1} + \frac{2}{\sqrt{3}}\arctan\frac{2x+1}{3} + C.$

2. $\int \frac{3x^4+2x^3+3x^2-1}{(x-2)(x^2+1)^2}dx = \int \frac{3}{x-2}dx + \int \frac{2\,dx}{x^2+1} + \int \frac{x\,dx}{(x^2+1)^2} =$
$= 3\ln|x-2| + 2\arctan x + \frac{1}{2}\int \frac{d(x^2+1)}{(x^2+1)^2} =$
$= 3\ln|x-2| + 2\arctan x - \frac{1}{2(x^2+1)} + C.$

3. $\int \frac{x+1}{(x-1)x(x-2)}dx = \int \frac{-2}{x-1}dx + \int \frac{dx}{2x} + \int \frac{3}{2(x-2)}dx =$
$= -2\ln|x-1| + \frac{1}{2}\ln|x| + \frac{3}{2}\ln|x-2| + C.$

7.9. OSTROGRADSKY'S METHOD

Ostrogradsky* proposed an ingenious method of separating the *rational part* of the integral of a proper rational fraction $P(x)/Q(x)$.

Analysing the form of the integrals of the four partial fractions (7.54), we come to the following conclusions:

(1) Integrals of the fractions of types I and III, whose denominators contain a binomial or, respectively, a first-degree trinomial *are irrational functions* (they are equal to the logarithm or to the arc tangent).

(2) The integral of the fraction of type II, whose denominator contains a binomial of degree $\beta > 1$, *is a proper rational fraction with a denominator which is equal to the same binomial of degree* $\beta - 1$.

(3) The integral of type IV, whose integrand function contains, in its denominator, a trinomial of degree λ, *is equal*, in a final analysis**, *to the sum of a proper rational fraction with a denominator equal to the same trinomial of degree* $\lambda - 1$, *and the integral* const $\int \frac{dx}{(x^2+px+q)}$ *which can be reduced to the arc tangent.*

Conclusions (1), (2), (3) enable us to make an inference as to the value of the rational part of the integral of the proper fraction $\frac{P(x)}{Q(x)}$, which we shall moreover consider to be *irreducible.*

---

* Academician M. V. Ostrogradsky (1801-1862), a Russian mathematician.

** With due account taken of the recurrence formula (6.12) derived at the end of 6.2.

Assume that the denominator $Q(x)$ has the form

$$Q(x) = (x-b_1)^{\beta_1} \dots (x-b_m)^{\beta_m} (x^2+p_1x+q_1)^{\lambda_1} \dots$$
$$\dots (x^2+p_nx-q_n)^{\lambda_n}. \tag{7.58}$$

Then the rational part of the integral of the proper rational fraction $\frac{P(x)}{Q(x)}$ is equal to the sum of proper rational fractions whose denominators are equal, respectively, to

$$(x-b_1)^{\beta_1-1}, \dots, (x-b_m)^{\beta_m-1}, \quad (x^2+p_1x+q_1)^{\lambda_1-1}, \dots$$
$$\dots, (x^2+p_nx+q_n)^{\lambda_n-1}.$$

The indicated sum* is, evidently, proper rational fraction $\frac{P_1(x)}{Q_1(x)}$ whose denominator $Q_1(x)$ has the form

$$Q_1(x) = (x-b_1)^{\beta_1-1} \dots (x-b_m)^{\beta_m-1} (x^2+p_1x+q_1)^{\lambda_1-1} \dots$$
$$\dots (x^2+p_nx+q_n)^{\lambda_n-1}. \tag{7.59}$$

Let us now calculate the sum of the partial fractions whose integrals are irrational functions. It follows from conclusions (1) and (3) that this sum is equal to the proper rational fraction $P_2(x)/Q_2(x)$, whose denominator $Q_2(x)$ is equal to

$$Q_2(x) = (x-b_1) \dots (x-b_m)(x^2+p_1x+q_1) \dots (x^2+p_nx+q_n). \tag{7.60}$$

We thus arrive at the following formula first derived by Ostrogradsky:

$$\int \frac{P(x)}{Q(x)}\,dx = \frac{P_1(x)}{Q_1(x)} + \int \frac{P_2(x)}{Q_2(x)}\,dx. \tag{7.61}$$

In Ostrogradsky's formula the polynomials $Q_1(x)$ and $Q_2(x)$ are specified by formulas (7.59) and (7.60) and can be calculated *without factorizing the polynomial $Q(x)$ into the product of irreducible factors.*

Indeed, in accordance with the results obtained in 7.4 (see formula (7.25)), the polynomial $Q_1(x)$ is the greatest common divisor of two polynomials $Q(x)$ and $Q'(x)$ and can be calculated with the aid of the Euclidean algorithm (see 7.4).

By virtue of formulas (7.58), (7.59) and (7.60), the polynomial $Q_2(x)$ is the quotient $\frac{Q(x)}{Q_1(x)}$ and can be calculated by means of the division of $Q(x)$ by $Q_1(x)$ "in a column".

It remains to calculate the polynomials $P_1(x)$ and $P_2(x)$. Since the

* That is the rational part of the integral of the fraction $\frac{P(x)}{Q(x)}$.

fractions $\frac{P_1(x)}{Q_1(x)}$ and $\frac{P_2(x)}{Q_2(x)}$ are proper, it is natural to give the polynomial $P_1(x)$ as a polynomial with undetermined coefficients of a degree lower by unity than that of $Q_1(x)$, and $P_2(x)$ is specified as a polynomial with undetermined coefficients of a degree lower by unity than that of $Q_2(x)$. To compute the indicated undetermined coefficients, we must differentiate Ostrogradsky's formula (7.61), reduce the results of the differentiation to a common denominator and compare the coefficients in the same degrees of x in the numerators.

Thus, Ostrogradsky's method is an ingenious technique of integrating a rational fraction without its preliminary decomposition into a sum of partial fractions. This technique is especially efficient in the case where the roots of $Q(x)$ are, in the main, multiple roots or where it is difficult to find the roots of $Q(x)$.

Example. Calculate by Ostrogradsky's method

$$\int \frac{6-7x-x^2}{x^4-2x^3+3x^2-2x+1}\,dx.$$

We have

$$Q(x) = x^4 - 2x^3 + 3x^2 - 2x + 1,$$
$$Q'(x) = 4x^3 - 6x^2 + 6x - 2.$$

We seek $Q_1(x)$ as the greatest common divisor of the polynomials $Q(x)$ and $Q'(x)$. Note that we have already found the greatest common divisor of *precisely these* two polynomials in the example considered at the end of 7.4. It is

$$Q_1(x) = x^2 - x + 1.$$

Divide $Q(x)$ by $Q_1(x)$ "in a column" and you will find

$$Q_2(x) = x^2 - x + 1.$$

$P_1(x)$ and $P_2(x)$ are given as first-degree polynomials with undetermined coefficients.

Ostrogradsky's formula (7.61) assumes the form

$$\int \frac{6-7x-x^2}{x^4-2x^3+3x^2-2x+1}\,dx = \frac{Ax+B}{x^2-x+1} + \int \frac{Cx+D}{x^2-x+1}\,dx. \tag{7.62}$$

To determine the coefficients A, B, C, D, we shall differentiate formula (7.62). We shall have

$$\frac{6-7x-x^2}{x^4-2x^3+3x^2-2x+1} = \frac{A(x^2-x+1)-(Ax+B)(2x-1)}{(x^2-x+1)^2} + \frac{Cx+D}{(x^2-x+1)}.$$

We reduce the result of differentiation to a common denominator and then compare the numerators. We receive

$$6 - 7x - x^2 = A(x^2 - x + 1) - (Ax + B)(2x - 1) + (Cx + D)(x^2 - x + 1).$$

Comparing the coefficients in x^0, x^1, x^2 and x^3, we obtain a system of equations

$$\left.\begin{aligned} &C = 0, \\ &-A + D - C = -1, \\ &-2B - D + C = -7, \\ &A + B + D = 6. \end{aligned}\right\}$$

Solving the system, we find that $A = 2$, $B = 3$, $C = 0$, $D = 1$. Thus, formula (7.62) assumes the form

$$\int \frac{6 - 7x - x^2}{x^4 - 2x^3 + 3x^2 - 2x + 1}\,dx = \frac{2x + 3}{x^2 - x + 1} + \int \frac{dx}{x^2 - x + 1}.$$

Integrating the right-hand side, we finally obtain

$$\int \frac{6 - 7x - x^2}{x^4 - 2x^3 + 3x^2 - 2x + 1}\,dx = \frac{2x + 3}{x^2 - x + 1} + \frac{2}{\sqrt{3}} \arctan \frac{2x - 1}{\sqrt{3}} + C.$$

7.10. INTEGRATING SOME IRRATIONAL AND TRANSCENDENTAL EXPRESSIONS

We have established in previous sections that the integral of any rational fraction is an elementary function. In this section we shall consider *some other classes of functions integrable in terms of elementary functions.* As a rule, we shall reduce, by means of a substitution, the integral of the function being discussed to an integral of a rational fraction. We shall speak of the indicated substitution as a substitution rationalizing the integral of the function being considered.

7.10.1. Integrating certain trigonometric expressions. Let us agree to designate, in what follows, any rational function of two arguments x and y* by the symbol $R(x, y)$.

---

* A rational function of two arguments is defined as follows. An nth degree polynomial of two arguments x and y is the expression of the form

$$P_n(x, y) = a_{00} + a_{10}x + a_{01}y + a_{20}x^2 + a_{11}xy + a_{02}y^2 + \ldots + a_{0n}y^n,$$

where $a_{00}, a_{10}, a_{01}, \ldots, a_{0n}$ are some constants. A rational function of two arguments is a ratio of the form $\frac{P_n(x, y)}{Q_m(x, y)}$, where $P_n(x, y)$ is an arbitrary polynomial of two arguments of degree n, and $Q_m(x, y)$ is an arbitrary polynomial of two arguments of degree m.

In this subsection we shall prove integrability, in terms of elementary functions, of any function of the form

$$R(\sin x, \cos x). \tag{7.63}$$

Let us prove that the integral of this function can be rationalized by means of the substitution $t = \tan\frac{x}{2}$. Indeed,

$$\sin x = \frac{2\tan\frac{x}{2}}{1+\tan^2\frac{x}{2}} = \frac{2t}{1+t^2}, \quad \cos x = \frac{1-\tan^2\frac{x}{2}}{1+\tan^2\frac{x}{2}} = \frac{1-t^2}{1+t^2},$$

$$x = 2\arctan t, \quad dx = \frac{2\,dt}{1+t^2},$$

so that

$$\int R(\sin x, \cos x)\,dx = \int R\left(\frac{2t}{1+t^2}, \frac{1-t^2}{1+t^2}\right)\frac{2\,dt}{1+t^2}.$$

Since a rational function of a rational function is also a rational function, the integral on the right-hand side of the last equation is an integral of a rational fraction.

Although $t = \tan\frac{x}{2}$ is a universal substitution rationalizing the integral of function (7.63), it often leads to cumbersome calculations. Therefore, we shall indicate some particular cases where the integral of function (7.63) can be rationalized with the aid of some other, more simple substitutions.

Notice, first of all, two elementary properties of a rational function of two arguments $R(u, v)$:

1°. If the rational function $R(u, v)$ does not change in value upon a change of sign of one of the arguments (say, u), that is, if $R(-u, v) = R(u, v)$, then this rational function can be reduced to the form $R(u, v) = R_1(u^2, v)$, where R_1 is some rational function of its two arguments. (This function contains only the *even* degrees of u.)

2°. Now if $R(u, v)$ changes sign upon a change of sign of u, i.e. $R(-u, v) = -R(u, v)$, it can be reduced to the form $R(u, v) = R_2(u^2, v)\,u$. (Property 2° follows immediately from Property 1° if it is applied to the function $\frac{R(u, v)}{u}$.)

Let us now take the problem of rationalizing the integral of function (7.63) for some particular cases.

I. Assume that $R(u, v)$ *changes sign upon a change of sign of* u. Then, in accordance with property 2°,

$$\int R(\sin x, \cos x)\,dx = \int R_2(\sin^2 x, \cos x)\sin x\,dx =$$

$$= -\int R_2(1-\cos^2 x, \cos x)\,d(\cos x).$$

Thus we see that the integral of the function (7.63) can be rationalized by means of the substitution $t = \cos x$.

II. Suppose now that the function $R(u, v)$ *changes sign upon a change of sign of* v. Then, in accordance with the same property 2°,

$$\int R(\sin x, \cos x)\,dx = \int R_3(\sin x, \cos^2 x)\cos x\,dx =$$

$$= \int R_3(\sin x, 1-\sin^2 x)\,d(\sin x),$$

that is, the integral of function (7.63) can be rationalized by means of the substitution $t = \sin x$.

III. Assume, finally, that the function $R(u, v)$ *does not change in value upon a simultaneous change of sign of* u *and* v, i.e.

$$R(-u, -v) = R(u, v).$$

Let us prove that in this case the integral of function (7.63) can be rationalized by means of the substitution $t = \tan x$. Indeed, in this case

$$R(u, v) = R\left(\frac{u}{v}v, v\right) = R_1\left(\frac{u}{v}, v\right),$$

$$R(-u, -v) = R\left(\frac{u}{v}(-v), -v\right) = R_1\left(\frac{u}{v}, -v\right).$$

Thus,

$$R_1\left(\frac{u}{v}, -v\right) = R_1\left(\frac{u}{v}, v\right).$$

But then, in accordance with property 1°,

$$R_1\left(\frac{u}{v}, v\right) = R_2\left(\frac{u}{v}, v^2\right).$$

We finally obtain

$$R(u, v) = R_2\left(\frac{u}{v}, v^2\right).$$

Hence

$$\int R(\sin x, \cos x)\,dx = \int R_2(\tan x, \cos^2 x)\,dx =$$

$$= \int R_2\left(\tan x, \frac{1}{1+\tan^2 x}\right)dx = \int R_2\left(t, \frac{1}{1+t^2}\right)\frac{dt}{1+t^2},$$

where $t = \tan x$, $x = \arctan t$, $dx = \dfrac{dt}{1+t^2}$.

Examples. (1) Calculate the integral $I_1 = \int \dfrac{dx}{1+a\cos x}$, where $a > 0$, $a \neq 1$. Applying the universal trigonometric substitution $t = \tan\dfrac{x}{2}$, we obtain

$$x = 2\arctan t, \quad dx = \frac{2\,dt}{1+t^2}, \quad \cos x = \frac{1-t^2}{1+t^2},$$

$$I_1 = 2\int \frac{dt}{(a+1)+t^2(1-a)} = \frac{2}{a+1}\int \frac{dt}{1+\frac{1-a}{1+a}t^2}.$$

Next we should separately consider two cases: (1) $0 < a < 1$, (2) $a > 1$.

In the case $0 < a < 1$

$$I_1 = \frac{2}{\sqrt{1-a^2}} \arctan\left(t\sqrt{\frac{1-a}{1+a}}\right) + C =$$

$$= \frac{2}{\sqrt{1-a^2}} \arctan\left(\sqrt{\frac{1-a}{1+a}} \tan\frac{x}{2}\right) + C.$$

In the case $a > 1$

$$I_1 = \frac{1}{\sqrt{a^2-1}} \ln\left|\frac{1+t\sqrt{\frac{a-1}{a+1}}}{1-t\sqrt{\frac{a-1}{a+1}}}\right| + C =$$

$$= \frac{1}{\sqrt{a^2-1}} \ln\left|\frac{1+\sqrt{\frac{a-1}{a+1}} \tan\frac{x}{2}}{1-\sqrt{\frac{a-1}{a+2}} \tan\frac{x}{2}}\right| + C.$$

(2) Calculate the integral $I_2 = \int \frac{\sin x\, dx}{\sin^2 x + 1}$.

Since the integrand function changes sign upon a change of sign of $\sin x$, there follows, in accordance with I, a substitution $t = \cos x$. The resulting expression will be

$$I_2 = -\int \frac{dt}{1-t^2+1} = \int \frac{dt}{t^2-2} = \frac{-1}{2\sqrt{2}} \ln\left|\frac{t+\sqrt{2}}{t-\sqrt{2}}\right| + C =$$

$$= -\frac{1}{2\sqrt{2}} \ln\left|\frac{\cos x+\sqrt{2}}{\cos x-\sqrt{2}}\right| + C.$$

(3) Calculate the integral $I_3 = \int \frac{\sin x \cos x}{\sin^4 x + \cos^4 x}\, dx$.

Since the integrand function retains its value upon a simultaneous change of sign of $\sin x$ and $\cos x$, we must make, in accordance with III, a substitution $t = \tan x$. The result will be

$$I_3 = \int \frac{t\, dt}{t^4+1} = \frac{1}{2} \int \frac{d(t^2)}{(t^2)^2+1} = \frac{1}{2} \arctan(t^2) + C = \frac{1}{2} \arctan(\tan^2 x) + C.$$

7.10.2. Integrating linear-fractional irrational functions. Here we shall prove integrability in elementary functions of any function of the form

$$R\left(x, \sqrt[n]{\frac{ax+b}{cx+d}}\right), \tag{7.64}$$

where a, b, c and d are some constants, n is any positive integer. Such a function will be referred to as a *linear-fractional irrational function.*

Let us prove that for $ad-bc\neq 0$ the integral of function (7.64) can be rationalized by means of substitution $t=\sqrt[n]{\frac{ax+b}{cx+d}}$. In fact,

$$t^n=\frac{ax+b}{cx+d},\quad x=\frac{dt^n-b}{a-c\cdot t^n},\quad dx=\frac{(ad-bc)\,nt^{n-1}}{(a-ct^n)^2}dt,$$

so that

$$\int R\left(x,\sqrt[n]{\frac{ax+b}{cx+d}}\right)dx=\int R\left(\frac{dt^n-b}{a-ct^n},t\right)\frac{(ad-bc)\,nt^{n-1}}{(a-ct^n)^2}dt.$$

Since a rational function of a rational function is also a rational function, the integral on the right-hand side of the last equation is an integral of a rational fraction. We have thus proved that the integral of the linear-fractional irrational function (7.64) can be rationalized by means of the substitution $t=\sqrt[n]{\frac{ax+b}{cx+d}}$.

Example. Calculate the integral $I=\int\sqrt{\frac{1+x}{1-x}}\,\frac{dx}{1-x}$. Having made the substitution

$$t=\sqrt{\frac{1+x}{1-x}},\quad t^2=\frac{1+x}{1-x},\quad x=\frac{t^2-1}{t^2+1},\quad dx=\frac{4t\,dt}{(t^2+1)^2},$$

we obtain

$$I=2\int\frac{t^2\,dt}{t^2+1}=2\int dt-2\int\frac{dt}{t^2+1}=2t-2\arctan t+C=$$

$$=2\sqrt{\frac{1+x}{1-x}}-2\arctan\sqrt{\frac{1+x}{1-x}}+C.$$

7.10.3. Integrating binomial differentials. A *binomial differential* is an expression of the form

$$x^m\,(a+bx^n)^p dx,$$

where a and b are any constants, and the exponents m, n and p are some rational numbers. Let us now discuss the problem of integrability of binomial differentials in terms of elementary functions.

First of all, we want to point out *three cases* where an integral of a binomial differential admits of a rationalizing substitution.

1°. The first case corresponds to an *integral* p. Here a binomial differential is a linear-fractional irrational function of the form $R\,(x,\sqrt[r]{x})\,dx$, where r is the least common multiple of the denominators of the rational numbers m and n. Consequently, in this case the integral of the binomial differential can be rationalized by means of the substitution $t=\sqrt[r]{x}$.

2°. The second case corresponds to the *integral number* $\frac{m+1}{n}$. Having made the substitution $z=x^n$ and having put, for brevity, $\frac{m+1}{n}-1=q$, we shall have

$$\int x^m\,(a+bx^n)^p\,dx=\frac{1}{n}\int(a+bz)^p\,z^q\,dz. \tag{7.65}$$

The integrand function on the right-hand side of (7.65) in a linear-fractional irrational function of the form $R\left(z, \sqrt[s]{a+bz}\right)$, where s is the denominator of the rational number p.

Thus, in the second case the binomial differential can be rationalized by means of the substitution

$$t=\sqrt[s]{a+bz}=\sqrt[s]{a+bx^n}.$$

3°. The third case corresponds to the *integral number* $\left(\frac{m+1}{n}+p\right)$. Here the integral function on the right-hand side of (7.65) is a linear-fractional irrational function of the form $R\left(z, \sqrt[s]{\frac{a+bz}{z}}\right)$ so that the integral of the binomial differential can be rationalized by means of the substitution

$$t=\sqrt[s]{\frac{a+bz}{z}}=\sqrt[s]{\frac{a}{x^n}+b}.$$

At the middle of the last century Chebyshev* proved that *the indicated three cases exhaust all the cases where a binomial differential can be integrated in terms of elementary functions.*

Examples. (1) Calculate the integral

$$I=\int \frac{dx}{x^2\sqrt{a+bx^2}}=\int x^{-2}\left(a+bx^2\right)^{-\frac{1}{2}}dx.$$

Here $m=-2$, $n=2$, $p=-\frac{1}{2}$, so that $\frac{m+1}{n}+p=-1$ (the third case). Having made the substitution

$$t=\sqrt{\frac{a}{x^2}+b}, \quad x=\frac{\sqrt{a}}{\sqrt{t^2-b}}, \quad dx=-\frac{\sqrt{a}\,t\,dt}{\sqrt{(t^2-b)^3}},$$

we obtain

$$I=\int\left(-\frac{dt}{a}\right)=-\frac{t}{a}+C=-\frac{\sqrt{\frac{a}{x^2}+b}}{a}+C.$$

(2) Calculate the integral $I=\int x^5\left(1-x^2\right)^{-\frac{1}{2}}dx$. In this case $m=5$, $n=2$, $p=-\frac{1}{2}$, so that $\frac{m+1}{n}=3$ (the second case). Having made the substitution

$$t=\sqrt{1-x^2}, \quad x=\sqrt{1-t^2}, \quad dx=-\frac{t\,dt}{\sqrt{1-t^2}},$$

we obtain

$$I=-\int(1-t^2)^2\,dt=-\int dt+2\int t^2\,dt-\int t^4\,dt=$$

$$=-t+\frac{2}{3}t^3-\frac{t^5}{5}+C=-\sqrt{1-x^2}+\frac{2}{3}\sqrt{(1-x^2)^3}-\frac{\sqrt{(1-x^2)^5}}{5}+C.$$

---

* P. L. Chebyshev (1821-1894), an eminent Russian mathematician.

7.10.4. Integrating quadratic irrational functions by means of Euler's substitution. Here we shall prove integrability in terms of elementary functions of any function of the form

$$R(x, \sqrt{ax^2+bx+c}), \tag{7.66}$$

where a, b and c are some constants. A function of this form will be referred to as a *quadratic irrational function.* In this case we are certain to assume that the quadratic trinomial $ax^2 + bx + c$ does not have *equal* roots (otherwise the root of this trinomial can be replaced by a rational expression).

We shall prove that the integral of function (7.66) can always be rationalized by one of *Euler's substitutions.*

We shall first consider the case where the quadratic trinomial $ax^2 + bx + c$ possesses *complex* roots. In this case the sign of the quadratic trinomial coincides with that of a and, since by the sense of our reasoning the quadratic trinomial (from which a quadratic root is being extracted) is *positive*, we have $a > 0$.

Thus, we have the right to make the following substitution:

$$t = \sqrt{ax^2+bx+c} + x\sqrt{a}. \tag{7.67}$$

Substitution (7.67) is usually called the *first Euler's substitution.* Let us prove that this substitution rationalizes the integral of function (7.66) for the case under consideration. Squaring both sides of the equality $\sqrt{ax^2 + bx + c} = t - x\sqrt{a}$, we obtain $bx + c = t^2 - 2\sqrt{a}\,tx$, so that

$$x = \frac{t^2-c}{2\sqrt{a}\,t+b}, \quad \sqrt{ax^2+bx+c} = \frac{\sqrt{a}\,t^2+bt+c\sqrt{a}}{2\sqrt{a}\,t+b},$$

$$dx = 2\,\frac{\sqrt{a}\,t^2+bt+c\sqrt{a}}{(2\sqrt{a}\,t+b)^2}\,dt.$$

Thus we have

$$\int R(x, \sqrt{ax^2+bx+c})\,dx =$$

$$= \int R\left(\frac{t^2-c}{2\sqrt{a}\,t+b}, \frac{\sqrt{a}\,t^2+bt+c\sqrt{a}}{2\sqrt{a}\,t+b}\right) 2\,\frac{\sqrt{a}\,t^2+bt+c\sqrt{a}}{(2\sqrt{a}\,t+b)^2}\,dt.$$

The integrand on the right-hand side of the equation is a rational fraction.

We shall now discuss the case where the quadratic trinomial $ax^2 + bx + c$ possesses *real* roots x_1 and x_2.

In this case we have $ax^2 + bx + c = a(x - x_1)(x - x_2)$. Let us prove that in this case the integral of function (7.66) can be ra-

tionalized by means of the substitution

$$t=\frac{\sqrt{ax^2+bx+c}}{x-x_1} \tag{7.68}$$

usually called the *second Euler's substitution*. Indeed, squaring the equality $\sqrt{ax^2+bx+c}=t(x-x_1)$ and cancelling the equality obtained by $(x-x_1)$, we receive $a(x-x_2)=t^2(x-x_1)$, so that

$$x=\frac{-ax_2+x_1t^2}{t^2-a},\quad \sqrt{ax^2+bx+c}=\frac{a(x_1-x_2)}{t^2-a}t,$$

$$dx=\frac{2a(x_1-x_2)t}{(t^2-a)^2}dt.$$

Thus we have

$$\int R(x,\sqrt{ax^2+bx+c})\,dx=$$

$$=\int R\left(\frac{-ax_2+x_1t^2}{t^2-a},\frac{a(x_1-x_2)t}{t^2-a}\right)\frac{2a(x_1-x_2)t}{(t^2-a)^2}dt.$$

The integrand on the right-hand side is a rational fraction.

Examples. (1) Calculate the integral $I=\int\frac{dx}{x+\sqrt{x^2+x+1}}$. Since the quadratic trinomial x^2+x+1 possesses complex roots, we shall make the first Euler's substitution

$$t=\sqrt{x^2+x+1}+x.$$

Squaring both sides of the equality $\sqrt{x^2+x+1}=t-x$, we obtain $x^2+x+1=t^2-2tx+x^2$ or $x+1=t^2-2tx$, so that

$$x=\frac{t^2-1}{1+2t},\quad dx=2\frac{t^2+t+1}{(1+2t)^2}dt.$$

Thus we have

$$I=2\int\frac{t^2+t+1}{t(1+2t)^2}dt=\int\left[\frac{A}{t}+\frac{B}{1+2t}+\frac{C}{(1+2t)^2}\right]dt.$$

Undetermined coefficients A, B and C can be easily calculated: $A=2$, $B=-3$, $C=-3$. We finally receive

$$I=2\ln|t|-\frac{3}{2}\ln|1+2t|+\frac{3}{2(1+2t)}+C=$$

$$=2\ln|\sqrt{x^2+x+1}+x|-\frac{3}{2}\ln|1+2x+2\sqrt{x^2+x+1}|+$$

$$+\frac{3}{2(1+2x+2\sqrt{x^2+x+1})}+C.$$

(2) Find the integral $I=\int\frac{dx}{1+\sqrt{1-2x-x^2}}$. Since the quadratic trinomial $1-2x-x^2$ has real roots $x_1=-1+\sqrt{2}$ and $x_2=-1-\sqrt{2}$,

we shall make the second Euler's substitution (7.68)

$$t=\frac{\sqrt{1-2x-x^2}}{x+1+\sqrt{2}}.$$

Squaring both sides of the equality $\sqrt{1-2x-x^2}=t(x+1+\sqrt{2})$ we shall have $(-1)(x+1-\sqrt{2})=t^2(x+1+\sqrt{2})$, so that

$$x=\frac{-t^2(\sqrt{2}+1)+\sqrt{2}-1}{t^2+1},\quad \sqrt{1-2x-x^2}=\frac{2\sqrt{2}}{t^2+1}t,$$

$$1+\sqrt{1-2x-x^2}=\frac{t^2+2\sqrt{2}t+1}{t^2+1},\quad dx=-\frac{4\sqrt{2}t}{(t^2+1)^2}dt.$$

Thus we have

$$I=-4\sqrt{2}\int\frac{t\,dt}{(t^2+1)(t^2+2\sqrt{2}t+1)}.$$

We have obtained an integral of a rational fraction whose calculation is left for the reader.

7.10.5. Integrating quadratic irrational functions by other means. Although Euler's substitutions can be always used to rationalize the integral of function (7.66), this substitution usually leads to cumbersome and complex calculations. In this connection, other, more practical, means are usually employed to integrate function (7.66). In what follows we shall be occupied with discussing these particular means.

Introducing the notation $y=\sqrt{ax^2+bx+c}$ and minding that y^2 is a polynomial, we can represent function (7.66) as the sum

$$R(x,\ y)=R_1(x)+\frac{R_2(x)}{y},$$

where $R_1(x)$ and $R_2(x)$ are some rational functions of one variable. Since $R_1(x)$ can be integrated (in terms of elementary functions), it suffices to evaluate the integral of the function $\frac{R_2(x)}{y}$.

We already know* that every rational fraction $R_2(x)$ can be represented as the sum of the polynomial $P(x)$ and the *proper* rational fraction $R_3(x)$. This latter can, in turn, be decomposed into the sum of partial fractions. Bearing this in mind, we can assert that the problem of integrating the function $\frac{R_2(x)}{y}$ reduces to finding the integrals of the following *three types*:

I. $\int\frac{P(x)}{y}dx$, where $P(x)$ is a polynomial.

II. $\int\frac{B}{(x-A)^\alpha y}dx$, where A and B are some constants, α is a natural number.

* See the beginning of Sec. 8.

III. $\int \frac{Mx+N}{(x^2+px+q)^\lambda y}\,dx$, where M, N, p and q are some constants, λ is a natural number, with $q-\frac{p^2}{4}>0$.

Let us discuss separately the integrals of types I, II and III.

I. To compute the *integrals of type I*, we must first establish the recurrence formula for the integral

$$I_m=\int \frac{x^m\,dx}{y}, \text{ where } m=0, 1, 2, \ldots$$

For this purpose we shall integrate the following identity, verifiable by means of differentiation, assuming $m \geqslant 1$:

$$(x^{m-1}y)' = ma\,\frac{x^m}{y}+\left(m-\frac{1}{2}\right)b\,\frac{x^{m-1}}{y}+(m-1)\,c\,\frac{x^{m-2}}{y}.$$

The integration of this identity leads to the equality

$$x^{m-1}y = maI_m+\left(m-\frac{1}{2}\right)bI_{m-1}+(m-1)\,cI_{m-2}. \tag{7.69}$$

Taking $m = 1$ in equality (7.69), we find

$$I_1=\frac{1}{a}\,y-\frac{b}{2a}\,I_0. \tag{7.70}$$

Assuming then $m = 2$ in equality (7.69) and making use of the calculated value of I_1 (i.e. formula (7.70)), we find

$$I_2=\frac{1}{4a^2}\,(2ax-3b)\,y+\frac{1}{8a^2}\,(3b^2-4ac)\,I_0.$$

Proceeding with analogous reasoning, we arrive at the following general formula:

$$I_m = P_{m-1}(x)\,y + c_mI_0, \tag{7.71}$$

where $P_{m-1}(x)$ is a polynomial of degree $m-1$, and c_m is a constant. If in the integral of type I $P(x)$ is a polynomial *of degree n*, the integral of type I will be equal to the sum of the integrals $I_0, I_1, \ldots, I_n$ with some constant factors (coefficients of the polynomial $P(x)$). Consequently, by virtue of equality (7.71), we shall finally obtain the following formula for the integral of type I:

$$\int \frac{P(x)}{y}\,dx = Q_{n-1}(x)\,y+C_0\int \frac{dx}{y}. \tag{7.72}$$

In this formula $Q_{n-1}(x)$ is a polynomial of degree $n-1$, and C_0 is a constant. To determine the polynomial $Q_{n-1}(x)$ and the constant C_0, use is usually made of the *method of undetermined coefficients*. The polynomial $Q_{n-1}(x)$ is written as a polynomial with literal coefficients

$$Q_{n-1}(x) = A_0+A_1x+\ldots+A_{n-1}x^{n-1}.$$

Differentiating equality (7.72) and multiplying the result by y_1, we obtain

$$P(x) = Q'_{n-1}(x)\,(ax^2+bx+c)+\frac{1}{2}\,Q_{n-1}(x)\,(2ax+b)+C_0. \tag{7.73}$$

Both sides of equality (7.73) contain polynomials of the nth degree. Equating their coefficients to each other, we obtain a system of $n+1$ linear equations, from which we can determine $A_0, A_1, \ldots, A_{n-1}$ and C_0.

The solvability of this system follows from the validity of formula (7.72) we have proved. It should only be added that the integral on the right-hand side of (7.72) can be reduced to the tabulated integral by means of a linear substitution of the variable $t = x + \frac{b}{2a}$. By means of the indicated substitution the integral $\int \frac{dx}{y}$ can be reduced, with an accuracy to within a constant factor, to one of the following two integrals:

$$\int \frac{dt}{\sqrt{t^2 \pm k^2}} = \ln | t + \sqrt{t^2 \pm k^2} | + C \quad (k = \text{const} > 0)$$

or

$$\int \frac{dt}{\sqrt{k^2 - t^2}} = \arcsin \frac{t}{k} + C.$$

Example. Find the integral

$$\int \frac{x^3}{\sqrt{1 + 2x - x^2}} dx.$$

For the integral being considered formula (7.72) has the form

$$\int \frac{x^3}{\sqrt{1 + 2x - x^2}} dx = (A_0 + A_1 x + A_2 x^2) \sqrt{1 + 2x - x^2} + C_0 \int \frac{dx}{\sqrt{1 + 2x - x^2}}. \tag{7.74}$$

Differentiating this formula and multiplying the result by $\sqrt{1 + 2x - x^2}$, we obtain

$$x^3 = (A_1 + 2A_2 x)(1 + 2x - x^2) + (A_0 + A_1 x + A_2 x^2)(1 - x) + C_0.$$

Comparing the coefficients in x^3, x^2, x^1, x^0 (on the right-hand and left-hand sides) we receive a system of equations

$$\left.\begin{aligned} -3A_2 &= 1, \\ 5A_2 - 2A_1 &= 0, \\ 2A_2 + 3A_1 - A_0 &= 0, \\ A_1 + A_0 + C_0 &= 0. \end{aligned}\right\}$$

Solving this system, we shall find $A_2 = -\frac{1}{3}$, $A_1 = -\frac{5}{6}$, $A_0 = -\frac{19}{6}$, $C_0 = 4$. The integral on the right-hand side of (7.74) can be computed by means of the substitution $t = x - 1$. We obtain

$$\int \frac{dx}{\sqrt{1 + 2x - x^2}} = \int \frac{dt}{2 - t^2} = \arcsin \frac{t}{\sqrt{2}} + C = \arcsin \frac{x - 1}{\sqrt{2}} + C.$$

The final result is

$$\int \frac{x^3}{\sqrt{1 + 2x - x^2}} dx = \left(-\frac{19}{6} - \frac{5}{6} x - \frac{1}{3} x^2 \right) \sqrt{1 + 2x - x^2} + 4 \arcsin \frac{x - 1}{\sqrt{2}} + C.$$

II. Let us now turn to the *integral of type II*. We shall show that this integral can be reduced to that of type I by the substitution $t = \frac{1}{x - A}$. Indeed, since

$$dx = -\frac{dt}{t^2}, \quad ax^2 + bx + c = \frac{(A^2a + Ab + c)\,t^2 + (2aA + b)\,t + a}{t^2}$$

we obtain

$$\int \frac{B}{(x-A)^\alpha y}\,dx = -\int \frac{Bt^{\alpha-1}\,dt}{\sqrt{(A^2a + Ab + c)\,t^2 + (2aA + b)\,t + a}}.$$

III. Let us now calculate the *integral of type III*. We shall begin with computing this integral *for the particular case* $p = 0$, $b = 0$, i.e. we shall compute the integral

$$K = \int \frac{Mx + N}{(x^2 + q)^\lambda \sqrt{ax^2 + c}}\,dx.$$

This integral decomposes into the sum of two integrals

$$K_1 = M \int \frac{x\,dx}{(x^2 + q)^\lambda \sqrt{ax^2 + c}} \text{ and } K_2 = N \int \frac{dx}{(x^2 + q)^\lambda \sqrt{ax^2 + c}}.$$

The first of these integrals can be written in the form

$$K_1 = \frac{M}{2} \int \frac{d(x^2)}{(x^2 + q)^\lambda \sqrt{ax^2 + c}},$$

which shows that the integrand function is a *linear* (and not quadratic) irrational function with respect to x^2. As was proved in 7.10.2, the integral K_1 can be rationalized by the substitution $t = \sqrt{ax^2 + c}$. The integral K_2 can be written in the form*

$$K_2 = N \int \frac{\frac{1}{x^{2\lambda-2}} \frac{dx}{x^3}}{\left(1 + q\frac{1}{x^2}\right)^\lambda \sqrt{a + c\frac{1}{x^2}}} =$$

$$= -\frac{N}{2} \int \frac{\left(\frac{1}{x^2}\right)^{\lambda-1} d\left(\frac{1}{x^2}\right)}{\left[1 + q\left(\frac{1}{x^2}\right)\right]^\lambda \sqrt{a + c\left(\frac{1}{x^2}\right)}},$$

which shows that the integrand function is a *linear* irrational function with respect to $\frac{1}{x^2}$. Consequently, the integral K_2 can be rationalized by means of the substitution $r = \sqrt{a + \frac{c}{x^2}}$. Thus, for the particular case where *none of the two quadratic trinomials contains first-degree terms*, the integral of type III has been rationalized.

Let us now consider the integral of type III *in the general case* and show that it can be reduced to the integral of the particular form discussed above. If the

* We assume that $x \neq 0$.

coefficients of quadratic trinomials satisfy the relation

$$b = ap, \tag{7.75}$$

then, to reduce the integral of type III to the integral of the particular form treated above, it is sufficient to make the substitution $x = t - \frac{p}{2}$. Indeed, having made the substitution, we shall obtain

$$\int \frac{(Mx+N)\,dx}{(x^2+px+q)^\lambda \sqrt{ax^2+bx+c}} =$$

$$= \int \frac{Mt+\left(\frac{Mp}{2}+N\right)}{\left[t^2+\left(q-\frac{p^2}{4}\right)\right]^\lambda \sqrt{at^2+\left(c-\frac{ap^2}{4}\right)}}\,dt.$$

It is much more difficult to reduce the integral of type III to the integral of the particular form discussed above for the case where the coefficients of the quadratic trinomials *do not satisfy* relation (7.75). In this case, we shall first make the linear-fractional substitution

$$x = \frac{\mu t+\nu}{1+t}, \tag{7.76}$$

having selected the constants μ and ν so that the quadratic trinomials obtained *do not contain any first-degree terms with respect to t.* We shall show that such μ and ν can indeed be selected. In fact, having made the substitution (7.76), we shall receive

$$x^2+px+q = \frac{(\mu^2+p\mu+q)\,t^2+[2\mu\nu+p\,(\mu+\nu)+2q]\,t+(\nu^2+p\nu+q)}{(1+t)^2},$$

$$ax^2+bx+c = \frac{(a\mu^2+b\mu+c)\,t^2+[2\mu\nu a+b\,(\mu+\nu)+2c]\,t+(a\nu^2+b\nu+c)}{(1+t)^2}.$$

Thus, the coefficients μ and ν can be determined from the system of equations

$$\left.\begin{aligned} 2\mu\nu+p\,(\mu+\nu)+2q=0,\\ 2\mu\nu a+b\,(\mu+\nu)+2c=0 \end{aligned}\right\}$$

or from the system of equivalent equations

$$\left.\begin{aligned} \mu+\nu &= -\frac{2\,(c-aq)}{b-ap},\\ \mu\cdot\nu &= \frac{cp-bq}{b-ap}. \end{aligned}\right\}$$

Thus we see that μ and ν are the roots of the quadratic equation

$$z^2+\frac{2\,(c-aq)}{b-ap}\,z+\frac{cp-bq}{b-ap}=0. \tag{7.77}$$

It remains to prove that quadratic equation (7.77) possesses real and distinct roots. It suffices to prove that the discriminant of this equation is positive, that is, to establish the inequality

$$(c-aq)^2 > (cp-bq)\,(b-ap). \tag{7.78}$$

It is easy to verify that inequality (7.78) is equivalent to the following inequality:

$$[2(c + aq) - bp]^2 > (4q - p^2)(4ac - b^2). \tag{7.79}$$

Since the quadratic trinomial $(x^2 + px + q)$ possesses complex roots, it follows that $4q - p^2 > 0$.

Inequality (7.79) obviously holds true in the case where $4ac - b^2 < 0$. Let us prove that this inequality also holds in the case of $4ac - b^2 > 0$. Here $q > 0$, $ac > 0$ and $4\sqrt{acq} > pb$. Therefore, since $\frac{c + aq}{2} \geqslant \sqrt{caq}$, we shall have

$$[2(c+aq) - bp]^2 \geqslant [4\sqrt{qac} - pb]^2 =$$

$$= (4q - p^2)(4ac - b^2) + 4(p\sqrt{ac} - b\sqrt{q})^2 \geqslant (4q - p^2)(4ac - b^2).$$

In the chain of inequalities written above there is at least one sign of strict inequality $>$, for the first sign $\geqslant$ turns into the sign $=$ only when $c = aq$, but for $c = aq$, by virtue of $b \neq ap$, the inequality $(p\sqrt{ac} - b\sqrt{q}) \neq 0$ obviously holds true and, therefore, the second sign $\geqslant$ does not turn into the sign $=$. Thus we have proved inequality (7.79), that is, we have proved the possibility of choosing μ and ν for which the obtained quadratic trinomials do not contain the first-degree terms with respect to t. Having made the substitution (7.76) with the indicated μ and ν, we shall reduce the integral of type III to the form

$$\int \frac{P(t)\,dt}{(t^2 + q_1)^{\lambda}\sqrt{a_1 t^2 + c_1}}, \tag{7.80}$$

where a_1, c_1 and q_1 are some constants, and $P(t)$ is a polynomial of degree $2\lambda - 1$. Having decomposed* the fraction $\frac{P(t)}{(t^2 + q_1)^{\lambda}}$ into the sum of partial fractions, we shall reduce the problem of finding integral (7.80) to the computation of the sum of the integrals of the form

$$\int \frac{M_k t + N_k}{(t^2 + q_1)^k \sqrt{a_1 t^2 + c_1}}\,dt \quad (k = 1, 2, \ldots, \lambda).$$

Each of these integrals is a particular integral of the type discussed above. Thus we have proved the integrability (in terms of elementary functions) of the integrals of all the three types I, II and III. And so we have once again proved the integrability of function (7.66) in terms of elementary functions without resorting to Euler's substitutions.

Example. Find the integral $I = \int \frac{dx}{(x^2 - x + 1)\sqrt{x^2 + x + 1}}$. This integral is of type III. Since for this integral relation (7.75) is violated, we must first of all make substitution (7.76). As a result we shall have

$$x^2 + x + 1 = \frac{(\mu^2 + \mu + 1)t^2 + [2\mu\nu + (\mu + \nu) + 2]t + (\nu^2 + \nu + 1)}{(1 + t)^2},$$

$$x^2 - x + 1 = \frac{(\mu^2 - \mu + 1)t^2 + [2\mu\nu - (\mu + \nu) + 2]t + (\nu^2 - \nu + 1)}{(1 + t)^2}.$$

The constants μ and ν can be found from the system of equations

$$2\mu\nu + (\mu + \nu) + 2 = 0,$$

$$2\mu\nu - (\mu + \nu) + 2 = 0.$$

* For $\lambda > 1$.

It is easy to verify that* $\mu = 1$, $\nu = -1$. Thus, substitution (7.76) has the form $x = \dfrac{t-1}{t+1}$, so that

$$t = \frac{x+1}{1-x}, \quad dx = \frac{2\,dt}{(1+t)^2}, \quad x^2+x+1 = \frac{3t^2+1}{(1+t)^2}, \quad x^2-x+1 = \frac{t^2+3}{(1+t)^2}.$$

The integral under consideration assumes the form

$$I = 2\int \frac{(1+t)\,dt}{(t^2+3)\sqrt{3t^2+1}} = I_1 + I_2,$$

where

$$I_1 = 2\int \frac{t\,dt}{(t^2+3)\sqrt{3t^2+1}}, \quad I_2 = 2\int \frac{dt}{(t^2+3)\sqrt{3t^2+1}}.$$

To compute the integral I_1, we make the substitution $u = \sqrt{3t^2+1}$, and to compute the integral I_2 we make the substitution $v = \sqrt{3+\frac{1}{t^2}}$. The result is

$$I_1 = 2\int \frac{du}{u^2+8} = \frac{1}{\sqrt{2}} \arctan \frac{u}{\sqrt{8}} + C = \frac{1}{\sqrt{2}} \arctan \sqrt{\frac{x^2+x+1}{2(1-x)^2}} + C,$$

$$I_2 = -2\int \frac{dv}{3v^2-8} = \frac{1}{2\sqrt{6}} \ln \left| \frac{v+\sqrt{\frac{8}{3}}}{v-\sqrt{\frac{8}{3}}} \right| + C =$$

$$= \frac{1}{2\sqrt{6}} \ln \left| \frac{\sqrt{\frac{x^2+x+1}{(x+1)^2}} + \sqrt{\frac{8}{3}}}{\sqrt{\frac{x^2+x+1}{(x+1)^2}} - \sqrt{\frac{8}{3}}} \right| + C.$$

7.11. ELLIPTIC INTEGRALS

Integrals of quadratic irrational functions are naturally related to the integrals

$$\int R(x, \sqrt{ax^3+bx^2+cx+d})\,dx \tag{7.81}$$

$$\int R(x, \sqrt{ax^4+bx^3+ex^2+dx+e})\,dx, \tag{7.82}$$

whose integrand functions contain a square root of the third- or fourth-degree polynomials.

These integrals are often encountered in applications. Notice first of all that integrals (7.81) and (7.82) *are not, generally speaking, elementary functions.*

It is usual to speak of these two integrals as of *elliptic* integrals in the cases where they are not expressed in terms of elementary functions and *pseudoelliptic* when they are expressed in terms of elementary functions**.

* We could have put $\mu = -1$, $\nu = 1$.

** The terms are due to the fact that these integrals were first encountered in problems of finding the circumference of an ellipse (see Example 3 in 11.1.6).

The importance of integrals (7.81) and (7.82) in applications evoked the necessity for tables and graphs of the functions determinable by these integrals. Since it is difficult to compile such tables and graphs for arbitrary coefficients a, b, c, d and e, a problem arose of reducing all integrals of types (7.81) and (7.82) to several types of integrals containing as few arbitrary coefficients as possible (or, as they say, of reducing integrals (7.81) and (7.82) to the *standard form*).

Note, first of all, that integral (7.81) can be reduced to form (7.82). Indeed, a cubic trinomial obviously has at least one real root x_0, and therefore it can be represented in the form $ax^3 + bx^2 + cx + d = a(x - x_0)(x^2 + px + q)$.

Having made the substitution $x - x_0 = \pm t^2$, we transform, as it is easy to see, integral (7.81) into (7.82).

Thus it is sufficient to consider integral (7.82) alone.

By virtue of the results of 7.6, a quartic polynomial can be factorized into the product of two quadratic trinomials with real coefficients

$$ax^4 + bx^3 + cx^2 + dx + e = a(x^2 + px + q)(x^2 + p'x + q').$$

There can always be chosen a certain linear or linear-fractional substitution eliminating linear terms in both quadratic trinomials*. Making such a substitution, we reduce integral (7.82) with an accuracy to within a summand which is an elementary function, to the form

$$\int \frac{R(t^2)\,dt}{\sqrt{A(1+mt^2)(1+m't^2)}}, \tag{7.83}$$

where R is an arbitrary rational function. Next, it is easy to show that with any combinations of absolute values and signs of the constants A, m and m', there can be found a substitution reducing integral (7.83) to the so-called *normal* (or *standard*) *integral*

$$\int \frac{R_1(z)^2\,dz}{\sqrt{(1-z^2)(1-k^2z^2)}}, \tag{7.84}$$

where k denotes a constant satisfying the condition $0 < k < 1$.

Any standard integral of type (7.84) can be reduced, with an accuracy to within a summand which is an elementary function, to the following three kinds:

$$\int \frac{dz}{\sqrt{(1-z^2)(1-k^2z^2)}}, \quad \int \frac{z^2\,dz}{\sqrt{(1-z^2)(1-k^2z^2)}} \tag{7.85}$$

and

$$\int \frac{dz}{(1+hz^2)\sqrt{(1-z^2)(1-k^2z^2)}} \quad (0 < k < 1).$$

Integrals (7.85) are called *elliptic integrals of the first, the second and the third kind* respectively. As was shown by Liouville**, each of these integrals is a *non-elementary function*. Elliptic integrals of the 1st and 2nd kinds contain only one parameter k assuming real values from the interval $0 < k < 1$, and the elliptic integral of the 3rd kind contains, in addition, a parameter h which can assume both real and complex values.

Legendre*** simplified integrals of the form (7.85) still further, having made the substitution $z = \sin\varphi \left(0 \leqslant \varphi \leqslant \frac{\pi}{2}\right)$.

---

* This fact can be proved in the same way as it is done in 7.10.5.

** Joseph Liouville (1809-1882), a French mathematician.

*** Adrien Marie Legendre (1752-1833), a French mathematician.

By means of this substitution, the first integral (7.85) can be reduced to the form

$$\int \frac{d\varphi}{\sqrt{1-k^2 \sin^2 \varphi}}. \tag{7.86}$$

The second integral (7.85) under the substitution with an accuracy to within a constant factor, turns out to be equal to the difference between integral (7.86) and the integral

$$\int \sqrt{1-k^2 \sin^2 \varphi}\, d\varphi. \tag{7.87}$$

The third integral (7.85) can be transformed into the integral

$$\int \frac{d\varphi}{(1+h \sin^2 \varphi)\sqrt{1-k^2 \sin \varphi}}. \tag{7.88}$$

Integrals (7.86), (7.87) and (7.88) are called *elliptic integrals of the first, the second, and the third kind, respectively, in Legendre's form.*

Integrals (7.86) and (7.87) play an especially important role in applications. If it is assumed that both these integrals vanish at $\varphi = 0$, then two quite definite functions are obtained which are usually denoted as $F(k, \varphi)$ and $E(k, \varphi)$. Extensive tables and graphs have been compiled for these functions. Legendre and other mathematicians studied the properties of these functions and derived a number of formulas for them.

Alongside elementary functions, the functions E and F have found their place in the class of functions often used in mathematical analysis. We should like to point out once again the conventionality of the notion of an elementary function. At the same time it should be emphasized that the study of functions integrable in terms of elementary functions is not the only aim of integral calculus.

CHAPTER 8

FUNDAMENTAL THEOREMS ON CONTINUOUS AND DIFFERENTIABLE FUNCTIONS

The concepts of a continuous and a differentiable function were introduced in Chapters 4 and 5. In this chapter we shall investigate a number of important properties of arbitrary continuous and differentiable functions. To establish these properties, we shall introduce a new definition of the limiting value of a function and prove the equivalence of this definition and that introduced in Chapter 4.

8.1. A NEW DEFINITION OF THE LIMITING VALUE OF A FUNCTION

8.1.1. A new definition of the limiting value of a function. Its equivalence to the definition given earlier. Let us assume, as we did in 4.2, that the function $y = f(x)$ is defined on a set $\{x\}$ and let a be an arbitrary point that does not necessarily belong to the set $\{x\}$ but possesses the property that in any ε-neighbourhood of point a there are points of the set $\{x\}$.

Recall the *old definition* of the limiting value of a function introduced in Chap. 4: *the number b is called the limiting value of the function $f(x)$ at point $x = a$ if for any, converging to a, sequence $x_1, x_2, \ldots, x_n, \ldots$ of the values of the argument x, whose terms differ from a, the corresponding sequence $f(x_1), f(x_2), \ldots, f(x_n), \ldots$ of the values of the function converges to b.*

Let us now formulate

A new definition of the limiting value of a function. *The number b is called the limiting value of the function $f(x)$ at point $x = a$ if for any positive number ε there can be found a positive number δ* such that for all the values of the argument x satisfying the inequality $0 < |x - a| < \delta$, the inequality $|f(x) - b| < \varepsilon$ holds true***.

Remark 1. The restriction $0 < |x - a|$ means that the values of the argument x being considered are *different from a*. This restric-

* Since δ depends on ε they sometimes write $\delta = \delta(\varepsilon)$.

** The old definition of the limiting value of a function is also called Heine's definition and the new definition is called Cauchy's definition of the limiting value.

tion becomes clear if we recall that the function $f(x)$ under study *may not be defined at point a*. Without this restriction it would be impossible to define the derivative $f'(a)$ as the limiting value of the function $F(x) = \frac{f(x)-f(a)}{x-a}$ at point a.

Remark 2. From the point of view of logic the main point in the new definition is that *for every* $\varepsilon > 0$ *there is a positive number* δ *corresponding to that* ε and ensuring the validity of the inequality

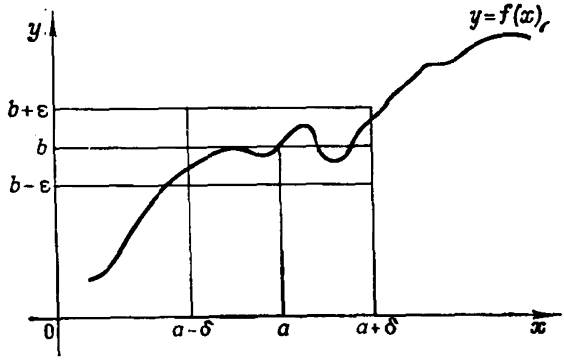


Fig. 8.1

$|f(x) - b| < \varepsilon$ for all the values of the argument x satisfying the inequality $0 < |x - a| < \delta$.

Remark 3. Making use of the idea of approximating the function $f(x)$ in the neighbourhood of point $x = a$ with the preassigned accuracy ε, we can reformulate the new definition of the limiting value of a function as follows: *the number b is called the limiting value of the function f (x) at point a if for any preassigned accuracy* ε *there can be chosen a* δ*-neighbourhood of point a such that for all the values of the argument x differing from a and belonging to the indicated* δ*-neighbourhood the number b approximates the value of the function f (x) with the accuracy* ε (Fig. 8.1).

Theorem 8.1. *The new definition of the limiting values of a function is equivalent to the old one.*

Proof. (1) First suppose that the number b is the limiting value of $f(x)$ at point a *in accordance with the new definition*. Let us prove that the same number b is the limiting value of $f(x)$ at point a *in accordance with the old definition* too. Assume that $\{x_n\}$ is any, converging to the number a, sequence of the values of the argument whose all terms are different from a. It is required to prove that the corresponding sequence $\{f(x_n)\}$ of the values of the function converges to the number b. Let us fix an *arbitrary* $\varepsilon > 0$. In accordance with the new definition of the limiting value of a function, there can be found, for this ε, $\delta > 0$ such that $|f(x) - b| < \varepsilon$ for all

the values of the argument x for which $0 < |x - a| < \delta$. Since the sequence $\{x_n\}$ converges to the number a, there can be chosen, for the indicated number $\delta > 0$, a number N such that $0 < |x_n - a| < \delta$ for $n \geqslant N$. Hence, $|f(x_n) - b| < \varepsilon$ for $n \geqslant N$, and this precisely means the convergence of the sequence $\{f(x_n)\}$ to the number b.

(2) Assume now that the number b is the limiting value of $f(x)$ at point a *in accordance with the old definition.* Let us prove that the same number b is the limiting value of $f(x)$ at point a *in accordance with the new definition too.* Suppose it is not so. Then, for some positive number ε *there cannot be chosen* a guaranteeing positive number δ indicated in the new definition, that is, for this ε and *for an arbitrarily small positive* δ there can be found at least one value of the argument x such that $0 < |x - a| < \delta$ but $|f(x) - b| \geqslant \varepsilon$.

By virtue of what was said above, we can take a sequence $\delta_n = \frac{1}{n}$ $(n = 1, 2, 3, \ldots)$ and assert that for its every term $\delta_n = \frac{1}{n}$ there can be found at least one value of the argument x_n such that

$$0 < |x_n - a| < \frac{1}{n}, \text{ but } |f(x_n) - b| \geqslant \varepsilon. \tag{8.1}$$

The left inequality (8.1) implies that the sequence $\{x_n\}$ converges to the number a and consists of terms different from a. But then, in accordance with the old definition of the limiting value of a function, the corresponding sequence $\{f(x_n)\}$ of the values of the function converges to the number b, but this is in contradiction with the right inequality (8.1) valid for all the numbers n. The contradiction obtained proves the theorem.

The new definition of the limiting value of a function enables us to formulate

A new definition of the continuity of a function at point $x = a$*.* *The function $f(x)$ is said to be continuous at point $x = a$ if for any positive number ε there can be found a positive number δ such that for all the values of the argument x satisfying the inequality $|x - a| < \delta$ there holds an inequality*

$$|f(x) - f(a)| < \varepsilon. \tag{8.2}$$

Remark 4. In this definition there is no need for the restriction $0 < |x - a|$, because at $x = a$ the left-hand side of inequality (8.2) vanishes and inequality (8.2) is obviously valid.

By analogy with the above statements a new definition of the limiting value of a function is formulated and the equivalence of

* In this case we assume, of course, that the function $y = f(x)$ is defined at point a itself as well.

this definition to the old one is proved *for the case where one of the numbers a and b or both of them turn into* $+\infty$ *or* $-\infty$. We shall formulate here a new definition of the limiting value of a function only for the case where $a = +\infty$: *the number b is called the limiting value of* $f(x)$ *for* $x \to +\infty$ *if for any positive number* ε *there can be found a positive number A such that for all the values of the argument x satisfying the inequality* $x > A$ *there holds an inequality* $|f(x) - b| < \varepsilon$.

Figure 8.2 is meant to explain this definition.

In conclusion we shall formulate a new definition of the *right-hand* and *left-hand* limiting values of the function $f(x)$ at point a: *the*

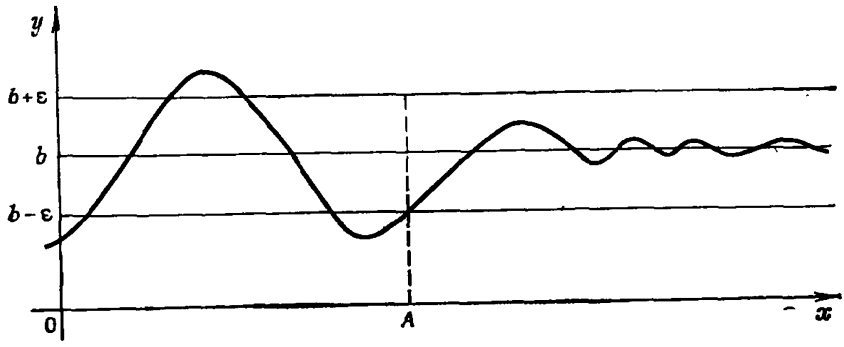


Fig. 8.2

number b is called the right-hand (left-hand) limiting value of the function $f(x)$ *at point a if for any positive number* ε *there can be found a positive number* δ *such that for all the values of the argument x satisfying the inequality* $0 < x - a < \delta$ $(0 < a - x < \delta)$, *the inequality* $|f(x) - b| < \varepsilon$ *holds true.*

The proof of the equivalence of this definition to the old definition of the right-hand (left-hand) limiting value is quite analogous to that of Theorem 8.1.

8.1.2. The necessary and sufficient condition of the existence of a limiting value (Cauchy's test). Because of the equivalence of the old and new definitions of the limiting value of a function, we can determine the necessary and sufficient condition of the existence of the limiting value of the function $f(x)$ at point a.

Definition. *We shall say that the function* $f(x)$ *satisfies Cauchy's test at point* $x = a$ *if for any positive number* ε *there can be found a positive number* δ *such that whatever the two values* x' *and* x'' *of the argument satisfying the inequalities* $0 < |x' - a| < \delta$, $0 < |x'' - a| < \delta$, *the inequality*

$$|f(x') - f(x'')| < \varepsilon$$

holds for the corresponding values of the function.

Theorem 8.2. (Cauchy's test). *For the function $f(x)$ to have a finite limiting value at point $x = a$, it is necessary and sufficient that the function $f(x)$ should satisfy Cauchy's test at that point.*

Proof. (1) **Necessity.** Suppose there exists a finite limiting value $\lim\limits_{x \to a} f(x) = b$. Let us prove that the function $f(x)$ satisfies Cauchy's test at point $x = a$. We shall take an arbitrary positive ε. In accordance with the new definition of the limiting value of a function, there can be found for the positive number $\varepsilon/2$ a positive number δ such that whatever the values x' and x'' of the argument satisfying the inequalities $0 < |x' - a| < \delta$, $0 < |x'' - a| < \delta$, the inequalities $|f(x') - b| < \varepsilon/2$, $|f(x'') - b| < \varepsilon/2$ hold true for the corresponding values of the function. Since the absolute value of the sum of two quantities does not exceed the sum of their absolute values, we have, from the last inequalities,

$$|f(x') - f(x'')| = |[f(x') - b] - [f(x'') - b]| \leqslant$$
$$\leqslant |f(x') - b| + |f(x'') - b| < \varepsilon.$$

Thus we have proved that the function $f(x)$ satisfies Cauchy's test at point $x = a$.

(2) **Sufficiency.** Suppose the function $f(x)$ satisfies Cauchy's test at point $x = a$. We shall prove that the function $f(x)$ has a limiting value at point $x = a$. Assume that $\{x_n\}$ is any sequence of the argument, converging to a, whose all terms x_n are different from a. By virtue of the old definition of the limiting value of a function it is sufficient to prove that the corresponding sequence $\{f(x_n)\}$ of the values of the function converges to some number b, this number b being *one and the same* for all the sequences $\{x_n\}$, converging to a, such that $x_n \neq a$.

We shall first prove the *convergence* of any sequences $\{f(x_n)\}$. Suppose we are given an *arbitrary* positive ε. We shall take a positive number δ corresponding to this ε in accordance with Cauchy's test and, making use of the convergence of the sequence $\{x_n\}$ to a, we shall choose for that δ a number N such that

$$0 < |x_n - a| < \delta \text{ for } n \geqslant N.$$

In this case, for any natural p ($p = 1, 2, \ldots$) the inequality

$$0 < |x_{n+p} - a| < \delta$$

is sure to hold for $n \geqslant N$. By virtue of Cauchy's test, the last two inequalities lead to the inequality $|f(x_{n+p}) - f(x_n)| < \varepsilon$ for $n \geqslant N$, that is they prove that the sequence $\{f(x_n)\}$ *is fundamental* (*Cauchy's sequence*). In accordance with Cauchy's test for a sequence (i.e. Theorem 3.19), the sequence $\{f(x_n)\}$ converges to a number b.

We shall now prove that all sequences $\{f(x_n)\}$ corresponding to various, converging to a, sequences $\{x_n\}$, *have one and the same limit* b.

Assume that $\{x_n\}$ and $\{x'_n\}$ are *any* two sequences of the values of the argument, converging to a, whose all terms are different from a. By virtue of what was proved above, both sequences $\{f(x_n)\}$ and $\{f(x'_n)\}$ are convergent. Let us designate the limit of the first sequence by b and the limit of the second sequence by b'. We shall prove that $b = b'$. Let us consider the sequence

$$x_1, x'_1, x_2, x'_2, \ldots, x_n, x'_n, \ldots$$

converging to a. As was proved above, the corresponding sequence of the values of the function

$$f(x_1), f(x'_1), f(x_2), f(x'_2), \ldots, f(x_n), f(x'_n), \ldots$$

is convergent. But then, by virtue of 3.4.1, *all the subsequences of that sequence*, the sequences $\{f(x_n)\}$ and $\{f(x'_n)\}$ among them, *converge to one and the same limit*, i.e. $b = b'$. We have proved Theorem 8.2.

The formulation of the Cauchy test and the establishment of the necessary and sufficient conditions for the existence of the limiting value of the function $f(x)$ at $x \to +\infty$ and at $x \to -\infty$ are quite analogous. We shall only give the formulations for the case of $x \to +\infty$.

We shall say that the function $f(x)$ *satisfies Cauchy's test at* $x \to +\infty$ *if for any positive number* ε *there is a positive number* A *such that for any two values of the argument* x' *and* x'' *exceeding* A *the inequality* $|f(x') - f(x'')| < \varepsilon$ *holds true.*

The following assertion can be proved by complete analogy with Theorem 8.2: *for the function* $f(x)$ *to have a finite limiting value at* $x \to +\infty$, *it is necessary and sufficient for it to satisfy the Cauchy test at* $x \to +\infty$.

8.2. A LOCAL BOUNDEDNESS OF A FUNCTION HAVING A LIMITING VALUE

By a complete analogy with the definition of a set of real numbers bounded above (below)* we shall introduce the concept of a *function bounded on the given set from above (from below)*.

Definition 1. *The function* $f(x)$ *is said to be* **bounded above (below)** *on the set* $\{x\}$, *if there is a real number* M *(number* m*) such that for all the values of the argument* x *from the set* $\{x\}$ *the inequality* $f(x) \leqslant M$ $(f(x) \geqslant m)$ *holds true.*

In this case the number M (the number m) is called the upper (the lower) bound of the function $f(x)$ on the set $\{x\}$.

* See 2.1.5.

Definition 2. *The function $f(x)$ is said to be bounded on both sides or simply* **bounded** *on the set $\{x\}$ if it is bounded on that set from above and from below, that is, if there are real numbers m and M such that for all the values of the argument x from the set $\{x\}$ the inequalities $m \leqslant f(x) \leqslant M$ hold true.*

Thus, the boundedness of the function $f(x)$ on the set $\{x\}$ actually means the boundedness of the set of all the values of that function.

Examples. (1) The function $f(x) = \sec x = \frac{1}{\cos x}$ on the half-open interval $\left[0, \frac{\pi}{2}\right)$ is not bounded above but is bounded below (any number $m \leqslant 1$ can be taken as the lower bound).

(2) The Dirichlet function* is bounded on any closed interval $[a, b]$ (any number $m \leqslant 0$ can serve as the lower bound and any number $M \geqslant 1$ as the upper bound).

Theorem 8.3. *If the function $f(x)$ possesses a finite limiting value at point $x = a$, then there exists a certain δ-neighbourhood of point a** such that for all the values of the argument from the indicated δ-neighbourhood the function $f(x)$ is bounded****.*

Proof. Assume $b = \lim\limits_{x \to a} f(x)$. In accordance with the new definition of the limiting value of a function, there can be chosen, for some positive number ε, a positive number δ such that $|f(x) - b| < \varepsilon$ as soon as $0 < |x - a| < \delta$ or

$$b - \varepsilon < f(x) < b + \varepsilon, \text{ as soon as } a - \delta < x < a + \delta \text{ and } x \neq a. \tag{8.3}$$

If the value $x = a$ *does not belong to the domain of definition of the function*, the theorem is proved (because inequalities (8.3) imply that *for all* the values of the argument x from the δ-neighbourhood of point a the values of the function $f(x)$ are contained between $b - \varepsilon$ and $b + \varepsilon$).

Now if the function $f(x)$ *is defined for* $x = a$ as well and assumes at point a a certain value $f(a)$, then, denoting by m the least of the two numbers $(b - \varepsilon)$ and $f(a)$ and by M the larger of the two numbers $(b + \varepsilon)$ and $f(a)$, we can obtain from inequalities (8.3) the following inequalities:

$$m \leqslant f(x) \leqslant M, \text{ as soon as } a - \delta < x < a + \delta.$$

---

* Recall that the Dirichlet function is a function equal to unity for all the rational values of the argument and to zero for all the irrational values of the argument.

** We wish to remind the reader that the δ-neighbourhood of point a is the interval $(a - \delta, a + \delta)$, where $\delta > 0$.

*** We do not exclude the case where the function $y = f(x)$ is specified on some set $\{x\}$, not filling entirely any δ-neighbourhood of point a.

The last inequalities imply that the function $f(x)$ is bounded everywhere in the δ-neighbourhood of point a. We have proved the theorem. Figure 8.3 serves as an illustration to Theorem 8.3.

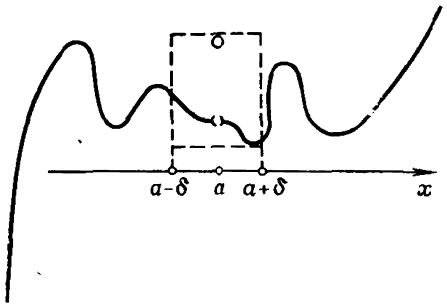


Fig. 8.3

Remark. The property of the function established by Theorem 8.3 is called the *local boundedness of the function possessing a limiting value.*

Corollary of Theorem 8.3. *If the function $f(x)$ is continuous at point $x = a$, this function is bounded for all the values of the argument from a certain δ-neighbourhood of point a.* (A function continuous at point $x = a$ possesses at that point a finite limiting value.)

8.3. THEOREM ON THE STABILITY OF SIGN OF A CONTINUOUS FUNCTION

Theorem 8.4. *If the function $f(x)$ is continuous at point $x = a$ and if $f(a) \neq 0$, there exists a δ-neighbourhood of point a such that for all the values of the argument from the indicated δ-neighbourhood the function $f(x)$ does not vanish and has a sign coinciding with that of $f(a)$.*

Proof. Since the function is continuous at point a, there exists $\lim\limits_{x \to a} f(x) = b$ with $b = f(a) \neq 0$. In accordance with the new definition of the limiting value of a function, there is $\delta > 0$ for *any* positive ε such that

$$b - \varepsilon < f(x) < b + \varepsilon \text{ as soon as}^* \ a - \delta < x < a + \delta. \quad (8.4)$$

Let us take for ε a positive number satisfying the requirement $\varepsilon < |b|$. With such a choice of ε the three numbers $b - \varepsilon$, $b + \varepsilon$ and b will be all *of the same sign.* Therefore, by virtue of (8.4), the function $f(x)$ retains the sign of the number $b = f(a)$ everywhere in the δ-neighbourhood of point a. The proof of the theorem is complete. Figure 8.4 can serve as an illustration to Theorem 8.4.

---

* There is no need in excluding the value $x = a$, since for the continuous function $f(x)$ the value $f(a) = b$ satisfies the left inequalities (8.4) as well.

Remark to Theorem 8.4. Theorem 8.4 can be extended to the case of the function $f(x)$ continuous at point $x = a$ *from the right (from the left)*. Let δ be a certain positive number. *We shall agree to refer to the half-open interval* $[a, a + \delta)$ *as the* **right-hand semi-neighbourhood** *of point* $x = a$ *and to the half-closed interval* $(a - \delta, a]$ *as the* **left-hand semi-neighbourhood** *of point* $x = a$. We can make the following statement: *if the function* $f(x)$ *is continuous at point* $x = a$ *from the right (from the left) and if* $f(a) \neq 0$, *then there is a right-hand (left-hand) semi-neighbourhood of point* $x = a$ *such that for all the values of the argument from the indicated semi-neighbourhood the function* $f(x)$ *does not vanish and is of the sign coinciding with that of* $f(a)$.

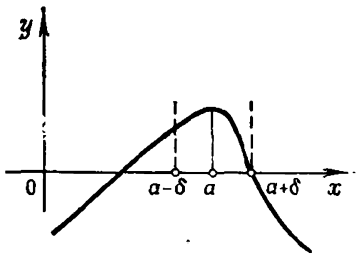


Fig. 8.4

The proof of this statement repeats almost verbatim the proof of Theorem 8.4, with the only exception that instead of the right-hand inequalities (8.4) we obtain inequalities $a \leqslant x < a + \delta$ $(a - \delta < x \leqslant a)$.

8.4. THE PASSAGE BY A CONTINUOUS FUNCTION THROUGH ANY INTERMEDIATE VALUE

8.4.1. The passage by a continuous function through zero at a change of signs.

Theorem 8.5. *Let the function* $f(x)$ *be continuous on the closed interval* $[a, b]$, *and let the values of this function at the ends* $f(a)$ *and* $f(b)$ *of the interval be some numbers of unlike signs. Then, there can be found, within the interval* $[a, b]$, *a point* ξ *at which the value of the function is equal to zero.*

Proof. Suppose, for the sake of definiteness, that $f(a) < 0$, $f(b) > 0$. Let us consider the set $\{x\}$ *of all the values of* x *belonging to the interval* $[a, b]$, *for which* $f(x) < 0$. This set possesses at least one member $x = a$ (since $f(a) < 0$) and is *bounded above* (say, by the value $x = b$). In accordance with Theorem 2.1, the set $\{x\}$ possesses a supremum which will be denoted by ξ.

Note, first of all, that point ξ is an *interior* point of the closed interval $[a, b]$, since the continuity of the function $f(x)$ on the interval $[a, b]$ and the conditions $f(a) < 0$, $f(b) > 0$ imply, by virtue of the remark to Theorem 8.4, that there is a right-hand semi-neighbourhood of point $x = a$ within which $f(x) < 0$, and a left-hand semi-neighbourhood of point $x = b$ within which $f(x) > 0$. Let us now prove that $f(\xi) = 0$. Were it not so, there would be found, in accordance with Theorem 8.4, a δ-neighbourhood $\xi - \delta < x < \xi + \delta$ of point ξ *within which the function* $f(x)$ *would be of*

a definite sign (either a plus sign everywhere or a minus sign everywhere). But that is impossible for, by the definition of the supremum, there is at least one value of x from the half-open interval $\xi - \delta < x \leqslant \xi$ such that $f(x) < 0$, and for any value of x belonging to the interval $\xi < x < \xi + \delta$ $f(x) \geqslant 0$. Thus we have $f(\xi) = 0$. We have proved the theorem. Figure 8.5 can serve as an illustration to Theorem 8.5.

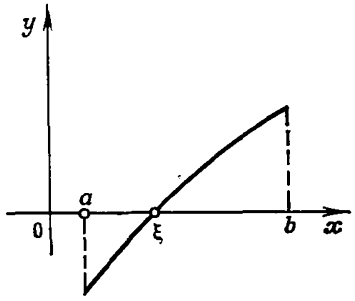


Fig. 8.5

8.4.2. The passage by a continuous function through any intermediate value.

Theorem 8.6. *Let the function $f(x)$ be continuous on the closed interval $[a, b]$, with $f(a) = A$ and $f(b) = B$. Assume, in addition, that C is any number contained between A and B. Then, there can be found, on the interval $[a, b]$ a point ξ such that $f(\xi) = C$.*

Proof. We need only consider the case where $A \neq B$ and where C does not coincide with any of the numbers A and B. Assume, for definiteness, that $A < B$, $A < C < B$. Let us consider the function $\varphi(x) = f(x) - C$. This function is continuous on the closed interval $[a, b]$ (as the difference between continuous functions) and assumes, at the end points of that interval, the values of unlike signs

$$\varphi(a) = f(a) - C = A - C < 0, \quad \varphi(b) = f(b) - C = B - C > 0.$$

In accordance with Theorem 8.5, there can be found, within the interval $[a, b]$, a point ξ such that $\varphi(\xi) = f(\xi) - C = 0$. Hence $f(\xi) = C$. We have proved the theorem.

8.5. THE BOUNDEDNESS OF A FUNCTION CONTINUOUS ON A CLOSED INTERVAL

Theorem 8.7 (Weierstrass' first theorem). *If the function $f(x)$ is continuous on the closed interval $[a, b]$, it is bounded on that interval.*

Proof. Let us prove that the function $f(x)$ is bounded *above* on the closed interval $[a, b]$ (the proof of boundedness from *below* is quite analogous).

Let us assume the contrary, i.e. suppose that $f(x)$ is not bounded above on the interval $[a, b]$.

Then, for *any* natural number n ($n = 1, 2, \ldots$), there is at least one point x_n, belonging to the interval $[a, b]$, such that $f(x_n) > n$ (otherwise $f(x)$ would be bounded above on the interval $[a, b]$).

Thus, there exists a sequence of the values x_n belonging to the interval $[a, b]$, such that the corresponding sequence of the values of the function $\{f(x_n)\}$ is infinitely large. By virtue of the Bolzano-

Weierstrass theorem (see Theorem 3.17 in 3.4.4), we can separate from the sequence $\{x_n\}$ a subsequence converging to a point ξ belonging to the interval $[a, b]$ by virtue of Remark 2 to the indicated theorem. Let us designate this subsequence as $\{x_{k_n}\}$ $(n = 1, 2, \ldots)$. By virtue of the continuity of the function $f(x)$ at point ξ the corresponding subsequence of the values of the function $\{f(x_{k_n})\}$ must necessarily converge to $f(\xi)$. But that is impossible, since the subsequence $\{f(x_{k_n})\}$ separated from the infinitely large sequence $\{f(x_n)\}$ is itself infinitely large (see 3.4.1). The contradiction obtained proves the theorem.

Remark. An assertion analogous to Theorem 8.7 is not valid for an open interval (or half-open interval), i.e. the continuity of a function on an open interval (or half-open interval) does not imply the boundedness of that function on the indicated set. Let us consider, as an example, the function $f(x) = 1/x$ on the interval $(0, 1)$ (or on the half-open interval $(0, 1]$). This function is continuous on the indicated interval (or the half-open interval) but is not bounded on it, for there exists a sequence of points $x_n = 1/n$ $(n = 2, 3, \ldots)$, belonging to the indicated interval (or the half-open interval), such that the corresponding sequence of the values of the function $\{f(x_n)\} = \{n\}$ is infinitely large.

8.6. THE LEAST UPPER BOUND AND THE GREATEST LOWER BOUND OF A FUNCTION AND THEIR ATTAINMENT BY A FUNCTION CONTINUOUS ON A CLOSED INTERVAL

8.6.1. The concepts of the least upper bound (the supremum) and the greatest lower bound (the infimum) of a function on a given set. Let us consider a function $f(x)$ bounded on a given set $\{x\}$ from above (from below)*. Making use of the concept of the least upper (greatest lower) bound, introduced in 2.1.5 for the set of all values of this function, we shall arrive at the following definition: *the number M (the number m) is referred to as the* **least upper (the greatest lower) bound of the function** $f(x)$ *on the set $\{x\}$ if the following two requirements are fulfilled*: (1) *the inequality $f(x) \leqslant M$ ($f(x) \geqslant m$) is valid for every value of x belonging to the set $\{x\}$*, (2) *whatever the positive number ε there is at least one value of x from the set $\{x\}$ for which the inequality*

$$f(x) > M - \varepsilon \quad (f(x) < m + \varepsilon)$$

is valid. In this definition requirement (1) asserts that the number M (number m) is one of the upper (lower) bounds of the function $f(x)$

* The definition of a function bounded on a given set from above (from below) was introduced at the beginning of Sec. 2 of this chapter.

on the set $\{x\}$, and requirement (2) says that this bound is the *least* (the *greatest*) and cannot be decreased (increased). The least upper bound and the greatest lower bound of the function $f(x)$ on the set $\{x\}$ are symbolized as follows:

$$M = \sup_{\{x\}} \{f(x)\}, \quad m = \inf_{\{x\}} \{f(x)\}.$$

Theorem 2.1 proved in 2.1.5 directly yields the following assertion: *if the function $f(x)$ is bounded on the set $\{x\}$ from above (from below), then the function $f(x)$ possesses on that set the least upper (the greatest lower) bound.*

The natural question arises of *whether the least upper (the greatest lower) bound is attainable*, i.e. whether there exists, among the points of the set $\{x\}$, a point x at which the value of the function is equal to that bound. The following example shows that the least upper bound and the greatest lower bound *are not, generally speaking, attainable.*

Let us consider a function

$$f(x) = \begin{cases} \sin x & \text{for } 0 < x < \frac{\pi}{2}, \\ \frac{1}{2} & \text{for } x = 0 \text{ and } x = \frac{\pi}{2}. \end{cases}$$

on the closed interval $[0, \pi/2]$. This function is bounded on the interval $[0, \pi/2]$ from above and from below and possesses on that interval the least upper bound $M = 1$ and the greatest lower bound $m = 0$. However, this function does not assume values equal to these bounds at any point of the interval $[0, \pi/2]$ (Fig. 8.6). Thus, the function being considered does not possess on the interval $[0, \pi/2]$ either the maximum or the minimum value.

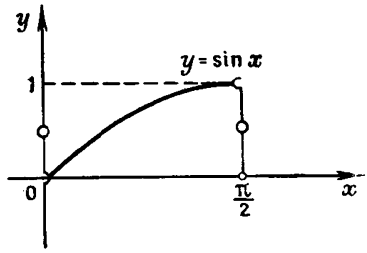


Fig. 8.6

Pay attention to the fact that the function we have considered is not continuous on the interval $[0, \pi/2]$. This is not accidental because, as will be proved in the next subsection, a function continuous on a closed interval necessarily attains its least upper bound and greatest lower bound at some points belonging to that interval.

8.6.2. Attaining by a function, continuous on a closed interval, its least upper and greatest lower bounds. Suppose a function $f(x)$ is continuous on a certain closed interval $[a, b]$. Then, by virtue of Theorem 8.7, this function is bounded on that interval both from above and from below. Therefore, in accordance with the definition formulated in the previous subsection, this function possesses on

the interval $[a, b]$ the least upper bound M and the greatest lower bound m. Let us prove that these bounds are attainable.

Theorem 8.8 (Weierstrass' second theorem). *If a function $f(x)$ is continuous on a closed interval $[a, b]$, then it attains on that interval its least upper and greatest lower bounds* (that is, there are, on the interval $[a, b]$ points x_1 and x_2 such that $f(x_1) = M$, $f(x_2) = m$).

Proof. Let us prove that the function $f(x)$ attains on the interval $[a, b]$ its least upper bound M (the attainment of the greatest lower bound is proved similarly).

Let us assume the contrary, that is, let us suppose that the function $f(x)$ does not take on a value equal to M at any point of the interval $[a, b]$. Then, the inequality $f(x) < M$ is valid *for all* the points belonging to the interval $[a, b]$ and we can consider a function

$$F(x) = \frac{1}{M - f(x)}$$

which is positive everywhere on the interval $[a, b]$. Since the denominator $M - f(x)$ does not vanish and is continuous on the interval $[a, b]$, the function $F(x)$ is also continuous on the interval $[a, b]$ in accordance with Theorem 4.2. In that case, according to Theorem 8.7, the function $F(x)$ is bounded on the interval $[a, b]$, that is, there is a positive number B such that for all x belonging to the interval $[a, b]$

$$F(x) = \frac{1}{M - f(x)} \leqslant B.$$

With due account of the fact that $M - f(x) > 0$, we can rewrite the last inequality in the form

$$f(x) \leqslant M - \frac{1}{B}.$$

The relation we have written, valid *for all* the points x belonging to the interval $[a, b]$, contradicts the fact that the number M is the supremum (*the least of all the upper bounds*) of the function $f(x)$ on the interval $[a, b]$. The contradiction obtained proves the theorem.

Remark 1. The assertion analogous to Theorem 8.8 does not hold for an open interval or a half-open interval. Indeed, in the remark to Theorem 8.7 (see 8.5) we cited an example of a function continuous on an open interval (a half-open interval) and not bounded on it (far from attaining its supremum and infimum, such a function, does not even possess them).

Remark 2. After we have proved that the function $f(x)$, continuous on a closed interval, attains on that interval its supremum and infimum, we can refer to the supremum as the *maximum value* and to the infimum as the *minimum value* of the function $f(x)$ on that interval and formulate Theorem 8.8 as follows: *a function*

*continuous on a closed interval possesses on that interval the maximum and the minimum value**.

Remark 3. Among the properties of a function continuous on a closed interval there is a property which can be referred to as a *uniform continuity*. This property will be discussed in 10.4 and, therefore, we want to point out that the whole text of 10.4.1 and 10.4.2 can be studied immediately after the material given in the present section.

8.7. A FUNCTION INCREASING (DECREASING) AT A POINT. A LOCAL MAXIMUM (MINIMUM)

8.7.1. An increase (decrease) of a function at a point. We shall suppose that the function $f(x)$ is defined everywhere in some neighbourhood of point c.

Definition. *The function $f(x)$ is said to* **increase (decrease) at point** c *if there is a neighbourhood of point throughout which $f(x) > f(c)$*

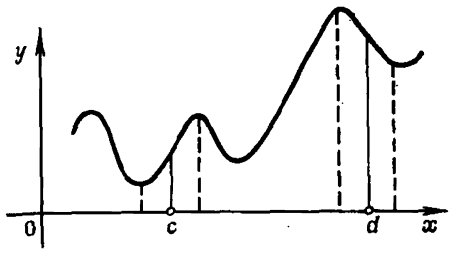


Fig. 8.7

for $x > c$ and $f(x) < f(c)$ for $x < c$ ($f(x) < f(c)$ for $x > c$ and $f(x) > f(c)$ for $x < c$).

Figure 8.7 illustrates a function increasing at point c and decreasing at point d.

We shall establish the *sufficient condition for an increase (decrease)* of the function $f(x)$ at point c.

Theorem 8.9. *If the function $f(x)$ is differentiable at point c and $f'(c) > 0$ ($f'(c) < 0$), then this function increases (decreases) at point c.*

---

* Make note of the fact that functions discontinuous on a certain closed interval can have on that interval the maximum and the minimum value. Thus, for example, the Dirichlet function

$$y = \begin{cases} 1, & \text{if } x \text{ is rational,} \\ 0, & \text{if } x \text{ is irrational} \end{cases}$$

familiar to us from 4.1, is discontinuous at any point of any closed interval $[a, b]$, but possesses on that interval the maximum value equal to unity and the minimum value equal to zero.

Proof. Let us prove the theorem for the case $f'(c) > 0$ (the case $f'(c) < 0$ can be proved by analogy). Since

$$f'(c) = \lim_{x \to c} \frac{f(x) - f(c)}{x - c},$$

the new definition of the limiting value of a function states that *for any* $\varepsilon > 0$ there can be found a positive δ such that

$$f'(c) - \varepsilon < \frac{f(x) - f(c)}{x - c} < f'(c) + \varepsilon \text{ for } 0 < |x - c| < \delta. \quad (8.5)$$

Let us take as ε a positive number smaller than $f'(c)$. Then $f'(c) - \varepsilon > 0$ and, consequently, we shall obtain from (8.5)

$$\frac{f(x) - f(c)}{x - c} > 0 \text{ for } 0 < |x - c| < \delta. \quad (8.6)$$

Relation (8.6) means that $f(x) > f(c)$ for $x > c$ and $f(x) < f(c)$ for $x < c$ *everywhere in the* δ*-neighbourhood of point* c. We have proved that the function $f(x)$ increases at point c.

Remark. We want to emphasize that *positivity (negativity) of the derivative* $f'(c)$ *is not a necessary condition for an increase (decrease) of the function* $f(x)$ *at point* c. We can cite the function $f(x) = x^3$ as an example which increases at point $x = 0$ and nevertheless possesses a derivative $f'(0) = 0$ at that point (the graph of this function is given in Fig. 8.8).

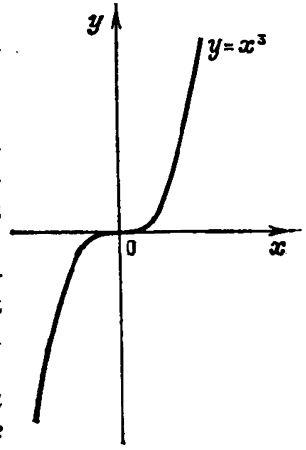


Fig. 8.8

8.7.2. A local maximum and a local minimum of a function. Assume once again that the function $f(x)$ is defined everywhere in a certain neighbourhood of point c.

Definition. *The function* $f(x)$ *is said to have a* **local maximum (minimum)** *at point* c *if there is a neighbourhood of point* c *within which the value* $f(c)$ *is the greatest (the least) among all the values of the function.*

Figure 8.9 illustrates the function $f(x)$ possessing a local maximum at point c.

The term **local extremum** is used to cover both a local maximum and a local minimum.

Let us establish a *necessary condition for an extremum* of a differentiable function.

Theorem 8.10. *If the function* $f(x)$ *is differentiable at point* c *and possesses a local extremum at that point, then* $f'(c) = 0$.

Proof. Since the function $f(x)$ has a local extremum at point c, the function $f(x)$ can neither increase nor decrease at that point. Therefore, by virtue of Theorem 8.9, the derivative $f'(c)$ can be neither positive nor negative, i.e. $f'(c) = 0$.

Theorem 8.10 has a simple geometrical meaning: it states that if at a point of a curve $y = f(x)$, at which the function $f(x)$ has a local extremum, there is a tangent to the graph of the function $y = f(x)$, then this tangent line is parallel to the x-axis (see Fig. 8.9).

8.8. THEOREM ON A ZERO OF A DERIVATIVE

Theorem 8.11 (Rolle's theorem*). *Suppose the function $f(x)$ is continuous on the closed interval $[a, b]$ and differentiable at all the interior points of that interval. Assume, in addition, that $f(a) = f(b)$.*

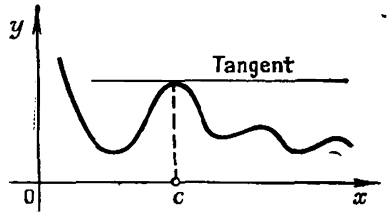


Fig. 8.9

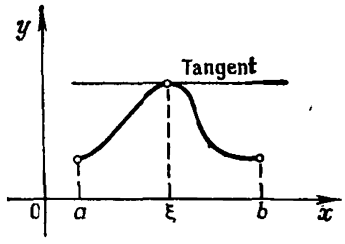


Fig. 8.10

Then, there will be a point ξ within the interval $[a, b]$ such that the value of the derivative $f'(\xi)$ at that point is equal to zero.

It can be said in short that between two equal values of a differentiable function there is necessarily a zero of the derivative of that function.

Proof. Since the function $f(x)$ is continuous on the closed interval $[a, b]$, then, according to Theorem 8.8, that function *reaches* its maximum value M and its minimum value m on this interval. Two cases may occur: (1) $M = m$, (2) $M > m$. In the first case $f(x) = M = m = \text{const}$. Therefore, the derivative $f'(x)$ *is equal to zero at any point of the interval* $[a, b]$. In the case $M > m$, inasmuch as $f(a) = f(b)$, we can assert that *at least one* of the values M and m is attained by the function at some *interior* point ξ of the interval $[a, b]$. But then the function $f(x)$ has a local extremum at that point ξ. Since the function $f(x)$ is differentiable at point ξ, $f'(\xi) = 0$ in accordance with Theorem 8.10. The proof of the theorem is complete.

Rolle's theorem has a simple geometrical meaning: if the extreme ordinates of the curve $y = f(x)$ are equal, then, in accordance with Rolle's theorem, there can be found, on the curve $y = f(x)$, a point at which a tangent to the curve is parallel to the x-axis (Fig. 8.10).

As will be seen later on, Rolle's theorem serves as the basis for many formulas and theorems of mathematical analysis.

* Michel Rolle (1652-1719), a French mathematician.

8.9. THE FORMULA OF FINITE INCREMENTS (LAGRANGE'S FORMULA)

The following theorem by Lagrange* is of great significance in analysis and its applications.

Theorem 8.12 (Lagrange's theorem). *If the function $f(x)$ is continuous on the closed interval $[a, b]$ and differentiable at all the interior points of that interval, then there can be found, within the interval $[a, b]$, a point ξ such that the formula*

$$f(b) - f(a) = f'(\xi)(b - a) \tag{8.7}$$

holds true.

Formula (8.7) is known as *Lagrange's formula* or the *formula of finite increments.*

Proof. Let us consider the following auxiliary formula on the closed interval $[a, b]$:

$$F(x) = f(x) - f(a) - \frac{f(b) - f(a)}{b - a}(x - a). \tag{8.8}$$

Let us verify that all the conditions of Rolle's theorem are fulfilled for the function $F(x)$. Indeed, $F(x)$ is continuous on the interval $[a, b]$ (as the difference between the function $f(x)$ and a linear function) and possesses a derivative equal to

$$F'(x) = f'(x) - \frac{f(b) - f(a)}{b - a}$$

at all the interior points of the interval $[a, b]$. It is evident from formula (8.8) that $F(a) = F(b) = 0$.

In accordance with Rolle's theorem, there is a point ξ within the closed interval $[a, b]$ such that

$$F'(\xi) = f'(\xi) - \frac{f(b) - f(a)}{b - a} = 0. \tag{8.9}$$

Equality (8.9) yields Lagrange's formula (8.7). We want to emphasize the fact that it is not necessary to assume $b > a$ in formula (8.7).

Remark. We have obtained Lagrange's theorem as a consequence of Rolle's theorem. But make note of the fact that Rolle's theorem is itself a particular case of Lagrange's theorem (for $f(a) = f(b)$).

To give Lagrange's theorem a geometric interpretation, we should point out that quantity $\frac{f(b) - f(a)}{b - a}$ is the slope of the secant passing through the points $A(a, f(a))$ and $B(b, f(b))$ of the curve $y = f(x)$, and $f'(\xi)$ is the slope of the tangent to the curve $y = f(x)$ passing through the point $C(\xi, f(\xi))$. Lagrange's formula

* Joseph Louis Lagrange (1736-1813), an eminent French mathematician and mechanician.

8.7) means that there is a point C on the curve $y=f(x)$ between the points A and B, at which the tangent is parallel to the secant AB (Fig. 8.11).

It is sometimes more convenient to write Lagrange's formula in a form somewhat different from (8.7). Suppose $f(x)$ satisfies the conditions of Theorem 8.11. We shall fix any x_0 from the closed interval $[a, b]$ and assign to it an increment Δx which is arbitrary but such that the value $(x_0 + \Delta x)$ necessarily lies on the interval $[a, b]$. Then, writing out Lagrange's formula for the interval $[x_0, x_0 + \Delta x]$, we shall have

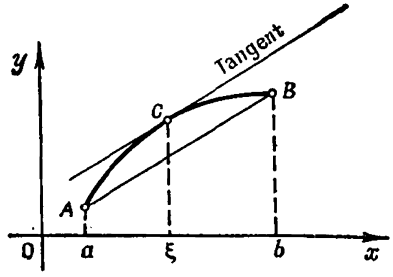


Fig. 8.11

$$f(x_0+\Delta x)-f(x_0)=\Delta x f'(\xi), \quad (8.10)$$

where ξ is some point contained between x_0 and $x_0 + \Delta x$. We can assert that *there is a number* (dependent on Δx) *belonging to the interval* $0 < \theta < 1$, such that

$$\xi = x_0 + \theta\Delta x.$$

Thus we can write formula (8.10) in the form

$$f(x_0+\Delta x)-f(x_0)=\Delta x f'(x_0+\theta\Delta x), \quad (8.11)$$

where θ is some number belonging to the interval $0 < \theta < 1$. Lagrange's formula in the form (8.11) gives a strict expression for the increment of the function through the arbitrary finite increment Δx of the argument which yielded it. This form of Lagrange's formula justifies the term "the formula of finite increments".

8.10. SOME COROLLARIES OF LAGRANGE'S FORMULA

8.10.1. The constant value of the function possessing a zero derivative on an interval.

Theorem 8.13. *If the function $f(x)$ is differentiable on the interval (a, b) and if $f'(x) = 0$ everywhere on that interval, then the function $f(x)$ is constant on the interval (a, b).*

Proof. Assume that x_0 is an arbitrary *fixed* point belonging to the interval (a, b) and x is *any* point of that interval.

The closed interval $[x_0, x]$ lies entirely on the interval (a, b). Therefore, the function $f(x)$ is differentiable (and consequently continuous) everywhere on the interval $[x_0, x]$. This fact makes it possible to apply Lagrange's theorem to the function $f(x)$ on the interval $[x_0, x]$. In accordance with this theorem, there is a point ξ within the interval $[x_0, x]$ such that

$$f(x) - f(x_0) = (x - x_0) f'(\xi). \quad (8.12)$$

By the hypothesis, the derivative of the function $f(x)$ is equal to zero everywhere on the interval (a, b). Hence $f'(\xi) = 0$, and we can obtain from formula (8.12)

$$f(x) = f(x_0). \tag{8.13}$$

Equality (8.13) states that the meaning of the function $f(x)$ *at any point* x of the interval (a, b) is equal to its value at the *fixed* point x_0. But this means that the function $f(x)$ is *constant everywhere on the interval* (a, b). The proof of the theorem is complete.

Theorem 8.13 has a simple geometrical meaning: if a tangent at each point of some portion of the curve $y = f(x)$ is parallel to the x-axis, then the indicated portion of the curve $y = f(x)$ is a segment of the straight line parallel to the x-axis.

Remark. We have already made use of Theorem 8.13 in Chapter 6 while proving Theorem 6.1. We want to point out here once again that we present the material in this chapter (including Theorem 8.13) without any resort to the results of Chapters 6 and 7. When you reread the book to study the material more thoroughly, you may read Chapter 8 immediately after Chapter 5, and then return to Chapters 6 and 7.

8.10.2. Conditions for monotonicity of a function on an interval. As the second corollary of Lagrange's formula, we shall discuss the conditions ensuring a nonincrease (nondecrease) of a function on a given open interval.

Let us recall, first of all, the definitions of nonincrease, nondecrease, increase and decrease of a function on a given open interval.

1°. The function $f(x)$ is said to be *nondecreasing* (*nonincreasing*) on an interval (a, b) if for any two points x_1 and x_2 of the interval (a, b) complying with the condition $x_1 < x_2$, the inequality

$$f(x_1) \leqslant f(x_2) \quad (f(x_1) \geqslant f(x_2))$$

holds true.

2°. The function $f(x)$ is said to be increasing (decreasing) on an interval (a, b) if for any two points x_1 and x_2 of the interval (a, b) connected by the condition $x_1 < x_2$, the inequality

$$f(x_1) < f(x_2) \quad (f(x_1) > f(x_2))$$

holds true.

Theorem 8.14. *For the function $f(x)$ differentiable on an interval (a, b) to be nondecreasing (nonincreasing) on that interval it is necessary and sufficient for the derivative of this function to be nonnegative (nonpositive) everywhere on that interval.*

Proof. (1) **Sufficiency.** Assume $f'(x) \geqslant 0$ ($\leqslant 0$) everywhere on the interval (a, b). It is required to prove that $f(x)$ does not decrease (does not increase) on the interval (a, b). Suppose that x_1 and x_2 are any two points belonging to the interval (a, b) and satisfying

the condition $x_1 < x_2$. The function $f(x)$ is differentiable (and, consequently, continuous) everywhere on the closed interval $[x_1, x_2]$. Therefore, Lagrange's theorem can be applied to the function $f(x)$ on the interval $[x_1, x_2]$. As a result we shall have

$$f(x_2) - f(x_1) = (x_2 - x_1) f'(\xi), \tag{8.14}$$

where $x_1 < \xi < x_2$.

By the hypothesis, $f'(\xi) \geqslant 0$ $(\leqslant 0)$, $x_2 - x_1 > 0$. Therefore, the right-hand side of (8.14) is nonnegative (nonpositive), and this proves a nondecrease (nonincrease) of $f(x)$ on the interval (a, b).

(2) **Necessity.** Assume that the function $f(x)$ is differentiable on the interval (a, b) and does not decrease (does not increase) on that interval. It is required to prove that $f'(x) \geqslant 0$ $(\leqslant 0)$ everywhere on the interval. Since $f(x)$ does not decrease (does not increase) on the interval (a, b), this function *cannot decrease (increase) at any point of the interval* (a, b). Hence, by Theorem 8.9, the derivative $f'(x)$ *cannot be negative (positive) at any point of the interval* (a, b). And that is what we had to prove.

Theorem 8.15. *For the function $f(x)$ to increase (decrease) on an interval (a, b) it is sufficient for its derivative $f'(x)$ to be positive (negative) everywhere on that interval.*

The proof is similar to that of the sufficiency in Theorem 8.14. Assume that x_1 and x_2 are any two points belonging to the interval (a, b) and satisfying the condition $x_1 < x_2$. Writing Lagrange's formula for the closed interval $[x_1, x_2]$, we shall obtain equality (8.14), but this time we shall have in this equality $f'(\xi) > 0$ (< 0).

Consequently, the left-hand side of (8.14) is positive (negative) which proves the increase (decrease) of $f(x)$ on the interval (a, b).

Remark. It should be emphasized that the positivity (negativity) of the derivative $f'(x)$ on the interval (a, b) *is not a necessary condition* of an increase (decrease) of the function $f(x)$ on the interval (a, b). For instance, the function $y = x^3$ increases on the interval $(-1, +1)$ but its derivative $f'(x) = 3x^2$ is not everywhere positive on that interval (it vanishes at point $x = 0$). In a general case, it is easy to prove that the function $f(x)$ increases (decreases) on the interval (a, b) if the derivative $f'(x)$ of this function is positive (negative) everywhere on that interval, *except for a finite number of points* at which the derivative is equal to zero. (To prove this statement, it is sufficient to apply Theorem 8.15 to each of the finite number of intervals on which $f'(x)$ is strictly positive (negative) and take into account the continuity of $f(x)$ at the points at which the derivative is equal to zero.) The connection between the sign of the derivative and the direction in which the function changes, established by Theorem 8.15, becomes clear if it is given geometrical interpretation. Since the derivative is equal to the slope of the tangent to the graph of the function $y = f(x)$, the sign of the derivative

shows whether the angle formed by the ray of the tangent lying in the upper half-plane and the positive direction of the x-axis is acute or obtuse. If $f'(x) > 0$ everywhere on the interval (a, b), then, everywhere on the interval, the ray of the tangent lying in the upper half-plane forms an acute angle with the x-axis and, hence, the curve $y = f(x)$ slants upward everywhere on that interval (Fig. 8.12).

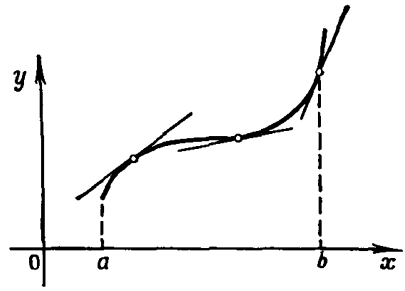


Fig. 8.12

8.10.3. A derivative possessing no points of discontinuity of the 1st kind or removable discontinuity. Let us apply Lagrange's theorem to elucidate one of the most remarkable properties of a derivative. First of all, let us prove the following assertion. *Postulate that the function $f(x)$ has a finite derivative everywhere in the right (left) semi-neighbourhood of point c and a right-hand (left-hand) derivative at point c itself. Then, if the derivative $f'(x)$ possesses at point c a right-hand (left-hand) limiting value, that limiting value is equal to the right-hand (left-hand) derivative at point c.*

To prove this assertion, we shall consider any sequence $\{x_n\}$ of the values of the argument converging to c from the right (from the left). Bearing in mind that, beginning with a sufficiently large n, all the x_n belong to the semi-neighbourhood in which the function $f(x)$ possesses a finite first derivative, we shall apply Lagrange's theorem to the function $f(x)$ with respect to the closed interval* $[c, x_n]$ $([x_n, c])$. We shall obtain

$$\frac{f(x_n)-f(c)}{x_n-c} = f'(\xi_n), \tag{8.15}$$

where ξ_n denotes a point lying between c and x_n. Suppose now $n \to \infty$ in equality (8.15). Then, evidently, $\xi_n \to c$ from the right (from the left). Since by the hypothesis $f'(x)$ possesses a finite right-hand (left-hand) limiting value at point c, the right-hand side of (8.15), by the definition of the limiting value, must necessarily tend to the indicated limit as $n \to \infty$. Hence, for $n \to \infty$ there exists a limit for the left-hand side of (8.15) as well. By the definition of the right-hand (left-hand) derivative this limit is equal to $f'(c+0)$ $(f'(c-0))$. Thus, in the limit, when $n \to \infty$, equality (8.15) yields

$$f'(c+0) = \lim_{x \to c+0} f'(x), \quad (f'(c-0) = \lim_{x \to c-0} f'(x)).$$

* All the conditions of Lagrange's theorem are fulfilled, for the function $f(x)$ is differentiable (and, hence, continuous) at any point of the closed interval $[c, x_n]$ $([x_n, c])$ except for point c. The continuity of $f(x)$ at point c from the right (from the left) follows from the existence of $f'(c+0)$ $(f'(c-0))$.

Thus we have proved that the function $f'(x)$ is continuous at point c from the right (from the left).

Making use of the statement just proved at each point c belonging to a certain interval (a, b), we shall arrive at the following assertion: *if the function $f(x)$ possesses a finite derivative everywhere on the interval (a, b), then $f'(x)$ cannot have on that interval either points of removable discontinuity or points of discontinuity of the 1st kind.*

Indeed, if at some point c belonging to the interval (a, b) the derivative $f'(x)$ possesses finite right-hand and left-hand limiting values, then $f'(x)$ *is continuous at point c* (by virtue of the assertion proved above). Now if even one of the indicated limiting values is absent, $f'(x)$ *has at point c a discontinuity of the 2nd kind.* Here is an example of a function whose derivative exists and is finite everywhere on a certain interval and possesses at some point of the interval a discontinuity of the 2nd kind. Let us consider a function

$$f(x)=\begin{cases} x^2\cos\frac{1}{x} & \text{for } x\neq 0,\\ 0 & \text{for } x=0 \end{cases}$$

on the interval $(-1, +1)$. It is evident that for any $x \neq 0$ the derivative of this function exists and is specified by the formula $f'(x) = 2x\cos(1/x) + \sin(1/x)$. The existence of the derivative $f'(0)$ at point $x = 0$ directly follows from the existence of the limiting value

$$\lim_{\Delta x\to 0}\frac{f(0+\Delta x)-f(0)}{\Delta x}=\lim_{\Delta x\to 0}\Delta x\cos\frac{1}{\Delta x}=0.$$

The derivative $f'(x)$ has neither a right-hand nor a left-hand limit at point $x = 0$ since at point $x = 0$ the summand $2x\cos(1/x)$ has a limit zero while the summand $\sin(1/x)$ does not possess at that point either a right-hand or a left-hand limiting value (see the example at the end of 4.8.1).

8.10.4. Derivation of certain inequalities. We will show in conclusion how some very useful inequalities can be derived with the aid of Lagrange's theorem. We shall establish, for example, the following two inequalities:

$$|\sin x_1 - \sin x_2| \leqslant |x_1 - x_2|, \tag{8.16}$$

$$|\arctan x_1 - \arctan x_2| \leqslant |x_1 - x_2|. \tag{8.17}$$

(Here x_1 and x_2 can be treated as any values of the argument.) To establish inequality (8.16), we shall apply Lagrange's theorem to the function $f(x) = \sin x$ with respect to the closed interval $[x_1, x_2]$. We obtain

$$\sin x_1 - \sin x_2 = (x_1 - x_2) f'(\xi). \tag{8.18}$$

Taking into account that $f'(\xi) = \cos \xi$ and that $|\cos \xi| \leqslant 1$ for any ξ, we shall obtain, passing to absolute values in (8.18), inequality (8.16).

To establish inequality (8.17), Lagrange's theorem should be applied to the function $f(x) = \arctan x$ with respect to the integral $[x_1, x_2]$, and it should be taken into account that $f'(\xi) = \frac{1}{1+\xi^2} \leqslant 1$.

8.11. A GENERALIZED FORMULA OF FINITE INCREMENTS (CAUCHY'S FORMULA)

In this section we shall prove a theorem by Cauchy, generalizing Lagrange's theorem formulated above.

Theorem 8.16 (Cauchy's theorem). *If each of the two functions $f(x)$ and $g(x)$ is continuous on the closed interval $[a, b]$ and differentiable at all the interior points of that interval, and if, in addition, the derivative $g'(x)$ is nonzero everywhere in the interior of the interval $[a, b]$, then there is a point ξ within that interval such that the formula*

$$\frac{f(b)-f(a)}{g(b)-g(a)} = \frac{f'(\xi)}{g'(\xi)} \tag{8.19}$$

holds true. Formula (8.19) is known as the *generalized formula of finite increments* or *Cauchy's formula.*

Proof. We shall prove first of all that $g(a) \neq g(b)$. In fact, were it not so, all the conditions of Theorem 8.11 (Rolle's theorem) would be fulfilled for the function $g(x)$ on the closed interval $[a, b]$, and, in accordance with this theorem, there would be a point ξ in the interior of the interval $[a, b]$ such that $g'(\xi) = 0$. The latter equality contradicts the condition of the theorem. Thus we have $g(a) \neq g(b)$ and can consider the following auxiliary function:

$$F(x) = f(x) - f(a) - \frac{f(b)-f(a)}{g(b)-g(a)}[g(x) - g(a)]. \tag{8.20}$$

By virtue of the requirements imposed on the functions $f(x)$ and $g(x)$, the function $F(x)$ is continuous on the interval $[a, b]$ and differentiable at all the interior points of the interval. Besides, it is evident that $F(a) = F(b) = 0$. Thus, all the conditions of Theorem 8.11 (Rolle's theorem) are fulfilled for $F(x)$. In accordance with this theorem, there is a point ξ within the interval $[a, b]$ such that

$$F'(\xi) = 0. \tag{8.21}$$

Bearing in mind that $F'(x) = f'(x) - \frac{f(b)-f(a)}{g(b)-g(a)} g'(x)$ and utilizing equality (8.21), we have

$$f'(\xi) - \frac{f(b)-f(a)}{g(b)-g(a)} g'(\xi) = 0. \tag{8.22}$$

With $g'(\xi) \neq 0$, equality (8.22) yields Cauchy's formula (8.19). The proof of the theorem is complete.

Remark 1. Lagrange's formula (8.7) is a particular case of Cauchy's formula (8.19) for $g(x) = x$.

Remark 2. It is not necessary to assume $b > a$ in formula (8.19).

8.12. EVALUATION OF INDETERMINATE FORMS (L'HOSPITAL'S RULE)

8.12.1. Evaluating indeterminate expressions of the form $\frac{0}{0}$. We shall say that the ratio of two functions $\frac{f(x)}{g(x)}$ is, for $x \to a$, an indeterminate expression of the form $\frac{0}{0}$ in the case where

$$\lim_{x \to a} f(x) = \lim_{x \to a} g(x) = 0.$$

To evaluate this indeterminate form means to calculate the limiting value $\lim\limits_{x \to a} \frac{f(x)}{g(x)}$ (provided the limiting value exists).

The following theorem provides the rule for evaluating the indeterminate form $\frac{0}{0}$.

Theorem 8.17 (L'Hospital's rule*). *Suppose two functions $f(x)$ and $g(x)$ are defined and differentiable everywhere in some neighbourhood of point a, with the exception, perhaps, of the point a itself. Also suppose that*

$$\lim_{x \to a} f(x) = \lim_{x \to a} g(x) = 0$$

*and the derivative $g'(x)$ is nonzero everywhere in the neighbourhood of point a indicated above. Then, if there exists a (finite or infinite) limiting value***

$$\lim_{x \to a} \frac{f'(x)}{g'(x)}, \tag{8.23}$$

---

* Guillaume François Antoine de L'Hospital (1661-1704), a French mathematician.

** Note that the limiting value (8.23) may not exist whereas the limit of the ratio of the functions $\lim\limits_{x \to a} \frac{f(x)}{g(x)}$ does exist. For example, we can take $a = 0$, $f(x) = x^2 \sin \frac{1}{x}$, $g(x) = \sin x$. Thus, L'Hospital's rule is not always valid.

there also exists a limiting value $\lim\limits_{x \to a} \frac{f(x)}{g(x)}$, *with the formula*

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \lim_{x \to a} \frac{f'(x)}{g'(x)} \tag{8.24}$$

being valid.

Theorem (8.17) yields the rule for evaluating indeterminate expressions of the form 0/0 reducing the calculation of the limiting value of the ratio of two functions to the calculation of the limiting value of the ratio of their derivatives.

Proof. Postulate that $\{x_n\}$ is an arbitrary sequence of the values of the argument converging to a and consisting of numbers distinct from a. We shall begin to consider this sequence *from the number* n, *starting from which all* x_n *belong to the neighbourhood of point* a *indicated in the formulation of the theorem.* We shall complete the definition of the functions $f(x)$ and $g(x)$ at point a, putting them equal to zero at that point. Then, evidently, $f(x)$ and $g(x)$ will be *continuous* throughout the closed interval $[a, x_n]$ and differentiable at all the interior points of that interval. Besides, $g'(x)$ is nonzero everywhere in the interior of the interval. Thus, all the conditions of Theorem 8.16 (Cauchy's theorem) are fulfilled for $f(x)$ and $g(x)$ on the closed interval $[a, x_n]$. In accordance with this theorem, there is a point ξ_n within the interval $[a, x_n]$ such that

$$\frac{f(x_n) - f(a)}{g(x_n) - g(a)} = \frac{f'(\xi_n)}{g'(\xi_n)}. \tag{8.25}$$

Taking into account that $f(a) = g(a) = 0$ according to our redefinition, we can rewrite formula (8.25) as follows:

$$\frac{f(x_n)}{g(x_n)} = \frac{f'(\xi_n)}{g'(\xi_n)}. \tag{8.26}$$

Postulate now that $n \to \infty$ in formula (8.26). Then, evidently, $\xi_n \to a$. Since we have assumed the existence of the limiting value (8.23), the right-hand side of (8.26), for $n \to \infty$, must necessarily tend to that limiting value. Therefore, there exists a limit, for $n \to \infty$, of the left-hand side of (8.26) as well. By the definition of the limiting value of a function this limit is equal to $\lim\limits_{x \to a} \frac{f(x)}{g(x)}$. Thus, in the limit, equality (8.26) passes into equality (8.24) as $n \to \infty$. We have proved the theorem.

Remark 1. If we add, to the conditions of Theorem (8.17), the requirement of the continuity of the derivatives $f'(x)$ and $g'(x)$ at point a, then, under the condition $g'(a) \neq 0$, formula (8.24) can be rewritten as

$$\lim_{x \to a} \frac{f(x)}{g(x)} = \frac{f'(a)}{g'(a)}. \tag{8.27}$$

Remark 2. If the derivatives $f'(x)$ and $g'(x)$ satisfy the same requirements as the functions $f(x)$ and $g(x)$ themselves, then L'Hospital's rule can be used repeatedly (i.e. the limiting value of the ratio of the first derivatives of the functions $f(x)$ and $g(x)$ can be replaced by the limiting value of the *ratio of the second derivatives* of those functions). Then we obtain

$$\lim_{x\to a}\frac{f(x)}{g(x)}=\lim_{x\to a}\frac{f'(x)}{g'(x)}=\lim_{x\to a}\frac{f''(x)}{g''(x)}.$$

Remark 3. Theorem 8.17 can be easily extended to the case where the argument x tends not to a finite but *to an infinite limit* $a=+\infty$ or $a=-\infty$. We think it sufficient to formulate Theorem 8.17 for the case of $a=+\infty$. *Suppose two functions $f(x)$ and $g(x)$ are defined and differentiable everywhere on the half-line $c<x<\infty$. Assume also that $\lim\limits_{x\to+\infty} f(x)=\lim\limits_{x\to+\infty} g(x)=0$ and the derivative $g'(x)$ is nonzero on the indicated half-line. Then, with the existence of the limiting $\lim\limits_{x\to+\infty}\frac{f'(x)}{g'(x)}$ there also exists a limiting value $\lim\limits_{x\to+\infty}\frac{f(x)}{g(x)}$, and the equality*

$$\lim_{x\to+\infty}\frac{f(x)}{g(x)}=\lim_{x\to+\infty}\frac{f'(x)}{g'(x)}$$

holds true.

Examples.

(1) $\lim\limits_{x\to0}\dfrac{1-\cos x}{x^2}=\lim\limits_{x\to0}\dfrac{\sin x}{2x}=\dfrac{1}{2}$.

(2) The following limiting value can be calculated by twice applying L'Hospital's rule:

$$\lim_{x\to0}\frac{x-\sin x}{x^3}=\lim_{x\to0}\frac{1-\cos x}{3x^2}=\lim_{x\to0}\frac{\sin x}{6x}=\frac{1}{6}.$$

(3) By thrice applying L'Hospital's rule we can calculate the limiting value

$$\lim_{x\to0}\frac{x^4}{x^2+2\cos x-2}=\lim_{x\to0}\frac{4x^3}{2x-2\sin x}=$$

$$=\lim_{x\to0}\frac{12x^2}{2-2\cos x}=\lim_{x\to0}\frac{24x}{2\sin x}=12.$$

8.12.2. Evaluating indeterminate expressions of the form $\frac{\infty}{\infty}$. The ratio of two functions $\frac{f(x)}{g(x)}$ is said to be an indeterminate

expression of the form $\frac{\infty}{\infty}$ for $x \to a$ if

$$\lim_{x \to a} f(x) = \lim_{x \to a} g(x) = \infty^*. \tag{8.28}$$

To evaluate indeterminate form, that is, to calculate the limiting value $\lim_{x \to a} \frac{f(x)}{g(x)}$, the following assertion which is quite analogous to Theorem 8.17, holds true: *if the requirement* $\lim_{x \to a} f(x) = \lim_{x \to a} g(x) = 0$ *in the formulation of Theorem* 8.17 *is replaced by condition* (8.28), *then Theorem* 8.17 *will remain valid.*

To prove the assertion, let us consider an *arbitrary* sequence $\{x_n\}$ of the values of the argument converging to a from the right (or from the left). Assume that x_m and x_n are any two terms of the sequence with sufficiently large numbers m and n satisfying the condition $n > m$.

Applying Cauchy's formula (8.19) with respect to the closed interval $[x_m, x_n]$, we can assert that there is a point ξ_{mn} on that interval such that

$$\frac{f(x_n) - f(x_m)}{g(x_n) - g(x_m)} = \frac{f(x_n)}{g(x_n)} \cdot \frac{1 - \frac{f(x_m)}{f(x_n)}}{1 - \frac{g(x_m)}{g(x_n)}} = \frac{f'(\xi_{mn})}{g'(\xi_{mn})},$$

whence we have

$$\frac{f(x_n)}{g(x_n)} = \frac{f'(\xi_{mn})}{g'(\xi_{mn})} \cdot \frac{1 - \frac{g(x_m)}{g(x_n)}}{1 - \frac{f(x_m)}{f(x_n)}}.$$

If there exists a limit $\lim_{x \to a} \frac{f'(x)}{g'(x)} = A$, then, *for any* $\varepsilon > 0$, we can fix the number m so large that for any $n > m$ the fraction $\frac{f'(\xi_{mn})}{g'(\xi_{mn})}$ will deviate from the number A by less than $\frac{\varepsilon}{2}$. Furthermore, taking into account (8.28), we can find, for the given fixed m, a number n_0 such that for $n \geqslant n_0$ the fraction

$$\frac{1 - \frac{g(x_m)}{g(x_n)}}{1 - \frac{f(x_m)}{f(x_n)}}$$

will deviate from unity by less than $\frac{\varepsilon}{|A| + \frac{\varepsilon}{2}}$. But then for $n \geqslant n_0$ the fraction $\frac{f(x_n)}{g(x_n)}$ will deviate from the number A by less than $\frac{\varepsilon}{2} + |A| \times$

* We can take $+\infty$ or $-\infty$ instead of ∞.

$\times \frac{\varepsilon}{|A|+\frac{\varepsilon}{2}} + \frac{\varepsilon}{2} \frac{\varepsilon}{|A|+\frac{\varepsilon}{2}} = \frac{3}{2}\varepsilon$. And this means that the limiting value $\lim\limits_{x \to a} \frac{f(x)}{g(x)}$ exists and is equal to A.

Examples. (1) $\lim\limits_{x \to 0+0} \sqrt{x} \ln x = \lim\limits_{x \to 0+0} \frac{\ln x}{x^{-\frac{1}{2}}} = \lim\limits_{x \to 0+0} \frac{\frac{1}{x}}{\left(-\frac{1}{2}\right) x^{-\frac{3}{2}}} =$

$= -2 \lim\limits_{x \to 0+0} \sqrt{x} = 0.$

(2) The limiting value

$$\lim_{x \to +\infty} \frac{x^n}{e^x} = \lim_{x \to +\infty} \frac{nx^{n-1}}{e^x} = \lim_{x \to +\infty} \frac{n(n-1)x^{n-2}}{e^x} = \ldots = \lim_{x \to +\infty} \frac{n!}{e^x} = 0$$

can be calculated by means of an n-fold application of L'Hospital's rule.

8.12.3. Evaluating indeterminate expressions of other forms. Besides the indeterminate expressions $\frac{0}{0}$ and $\frac{\infty}{\infty}$, other indeterminate expressions of the form $0 \cdot \infty$, $\infty - \infty$, 1^∞, 0^0, ∞^0 are also encountered in practice.

All these indeterminate forms can be reduced to the two forms studied above by means of algebraic transformations. We will show this on the example of the *last three* indeterminate forms indicated above. Each of these indeterminate expressions has the form

$$y = f(x)^{g(x)}, \tag{8.29}$$

where, for $x \to a$, $f(x)$ tends to 1, 0 or ∞ respectively, and $g(x)$ tends to ∞, 0 or 0 respectively. Taking the logarithm of expression (8.29), we obtain (assuming that $f(x) > 0$)

$$\ln y = g(x) \ln f(x). \tag{8.30}$$

To find the limiting value of expression (8.29), it is sufficient to find the limiting value of expression (8.30).

Make note of the fact that in any of the three considered cases expression (8.30) is, for $x \to \infty$, an indeterminate expression of the form $0 \cdot \infty$. So it is sufficient to learn how to reduce indeterminate expressions of the form $0 \cdot \infty$ to the indeterminate expression of the form $\frac{0}{0}$ or $\frac{\infty}{\infty}$. We will now show how to do it. Postulate that

$$z = \varphi(x) \cdot \psi(x), \tag{8.31}$$

with

$$\lim_{\to a} \varphi(x) = 0, \quad \lim_{x \to a} \psi(x) = \pm \infty.$$

Let us rewrite (8.31) in the form

$$z=\varphi(x)\cdot\psi(x)=\frac{\varphi(x)}{\frac{1}{\psi(x)}}. \tag{8.32}$$

For $x\to a$ expression (8.32) is, evidently, an indeterminate expression of the form $\frac{0}{0}$. We have achieved our aim.

Examples. (1) Calculate $\lim\limits_{x\to 0+0} x^x$. We shall use the notation $y=x^x$. Then $\ln y=x\ln x=\frac{\ln x}{\frac{1}{x}}$. Applying L'Hospital's rule, we shall have

$$\lim_{x\to 0+0}(\ln y)=\lim_{x\to 0+0}\frac{\ln x}{\frac{1}{x}}=\lim_{x\to 0+0}\frac{\frac{1}{x}}{-\frac{1}{x^2}}=-\lim_{x\to 0+0}x=0.$$

Hence it is evident that $\lim\limits_{x\to 0+0} y=1$.

(2) $\lim\limits_{x\to 0}(1+x^2)^{\frac{1}{e^x-1-x}}$. Assume that $y=(1+x^2)^{\frac{1}{e^x-1-x}}$. Then

$$\ln y=\frac{1}{(e^x-1-x)}\cdot\ln(1+x^2).$$

Making use of L'Hospital's rule, we shall receive

$$\lim_{x\to 0}\ln y=\lim_{x\to 0}\frac{\ln(1+x^2)}{e^x-1-x}=\lim_{x\to 0}\frac{\frac{2x}{1+x^2}}{e^x-1}=$$

$$=\lim_{x\to 0}\frac{2x}{(e^x-1)(1+x^2)}=\lim_{x\to 0}\frac{2}{e^x(1+x^2)+(e^x-1)2x}=2.$$

It is clear now that $\lim\limits_{x\to 0} y=e^2$.

8.13. TAYLOR'S FORMULA

The formula we are going to establish in this section is one of the basic formulas of mathematical analysis which has numerous applications in the analysis itself and in the adjacent fields of mathematical science.

Theorem 8.18 (Taylor's* theorem). *Assume that the function $f(x)$ possesses a derivative of the order** $n+1$ (n being any fixed number)*

* Brook Taylor (1685-1731), a British mathematician.

** It follows from this that *the function $f(x)$ itself and its derivatives up to order n are continuous in the indicated neighbourhood of point a.*

in a certain neighbourhood of point a. Suppose, furthermore, that x is any value of the argument from the indicated neighbourhood and p is an arbitrary positive integer. Then there is a point ξ contained between the points a and x such that the formula

$$f(x)=f(a)+\frac{f'(a)}{1!}(x-a)+\frac{f^{(2)}(a)}{2!}(x-a)^2+\ldots$$

$$\ldots+\frac{f^{(n)}(a)}{n!}(x-a)^n+R_{n+1}(x), \tag{8.33}$$

holds true, in which*

$$R_{n+1}(x)=\left(\frac{x-a}{x-\xi}\right)^p\cdot\frac{(x-\xi)^{n+1}}{n!p}f^{(n+1)}(\xi). \tag{8.34}$$

Formula (8.33) is known as *Taylor's formula* (with centre at a) and the expression $R_{n+1}(x)$ is called the *remainder*. As we shall see later, the remainder can be written not only in form (8.34) but in other forms as well. The remainder written in form (8.34) is usually called a *remainder in a generalized form***.

Proof. Let us symbolize as $\varphi(x, a)$ a polynomial with respect to x of order n entering into the right-hand side of (8.33), i.e. let us put

$$\varphi(x, a)=f(a)+\frac{f'(a)}{1!}(x-a)+$$

$$+\frac{f^{(2)}(a)}{2!}(x-a)^2+\ldots+\frac{f^{(n)}(a)}{n!}(x-a)^n. \tag{8.35}$$

By the symbol $R_{n+1}(x)$ we shall denote the difference

$$R_{n+1}(x)=f(x)-\varphi(x, a). \tag{8.36}$$

The theorem will be proved if we determine that $R_{n+1}(x)$ is specified by formula (8.34).

Let us fix any value of x from the neighbourhood indicated in the formulation of the theorem. We shall assume for definiteness that $x>a$. We shall denote by t a variable whose range of variation is the closed interval $[a, x]$ and consider the auxiliary function $\psi(t)$ of the form

$$\psi(t)=f(x)-\varphi(x, t)-(x-t)^p Q(x), \tag{8.37}$$

where

$$Q(x)=\frac{R_{n+1}(x)}{(x-a)^p}. \tag{8.38}$$

---

* Since ξ lies between x and a, we have $\frac{x-a}{x-\xi}>0$, so that expression $\left(\frac{x-a}{x-\xi}\right)^p$ is defined for any positive p.

** This form of the remainder is also called the form of Schlömilch-Rouché.

In a more extensive form $\psi(t)$ can be written out as

$$\psi(t) = f(x) - f(t) - \frac{f'(t)}{1!}(x-t) - \ldots$$

$$\ldots - \frac{f^{(n)}(t)}{n!}(x-t)^n - (x-t)^p Q(x). \tag{8.39}$$

Our aim is to express $Q(x)$ proceeding from the properties of the function $\psi(t)$ we have introduced.

Let us show that the function $\psi(t)$ satisfies all the conditions of Theorem 8.11 (Rolle's theorem) on the closed interval $[a, x]$.

It is evident from formula (8.39) and from the conditions imposed on the function $f(x)$ that the function $\psi(t)$ is continuous on the interval $[a, x]$ and differentiable on that interval*. Let us make sure that $\psi(a) = \psi(x) = 0$. Putting $t = a$ in (8.37) and taking into account equality (8.38), we obtain

$$\psi(a) = f(x) - \varphi(x, a) - R_{n+1}(x),$$

whence, on the basis of (8.36), we obtain $\psi(a) = 0$. The equality $\psi(x) = 0$ follows immediately from formula (8.39).

Thus, all the conditions of Theorem 8.11 (Rolle's theorem) are fulfilled for the function $\psi(t)$ on the interval $[a, x]$. On the basis of this theorem, there can be found, within the interval $[a, x]$, a point ξ such that

$$\psi'(\xi) = 0 \tag{8.40}$$

Let us calculate the derivative $\psi'(t)$. Having differentiated equality (8.39), we obtain

$$\psi'(t) = -f'(t) + \frac{f'(t)}{1!} - \frac{f^{(2)}(t)}{1!}(x-t) + \frac{f^{(2)}(t)}{2!}2(x-t) - \ldots$$

$$\ldots + \frac{f^n(t)}{n!}n(x-t)^{n-1} - \frac{f^{(n+1)}(t)}{n!}(x-t)^n + p(x-t)^{p-1}Q(x). \tag{8.41}$$

It is easy to see that all the terms on the right-hand side of (8.41), with the exception of the last two terms, can be cancelled out. Thus,

$$\psi'(t) = -\frac{f^{(n+1)}(t)}{n!}(x-t)^n + p(x-t)^{p-1}Q(x). \tag{8.42}$$

Putting $t = \xi$ in formula (8.42) and making use of equality (8.40), we receive

$$Q(x) = \frac{(x-\xi)^{n-p+1}}{n!p} f^{(n+1)}(\xi). \tag{8.43}$$

* For $f(t)$ and its derivatives up to the nth order are continuous on the closed interval $[a, x]$ and $f^{(n+1)}(t)$ exists and is finite on that interval (see footnote** on p. 269).

Comparing (8.43) and (8.38), we shall finally have

$$R_{n+1}(x) = (x-a)^p Q(x) = \frac{(x-a)^p (x-\xi)^{n-p+1}}{n!p} f^{(n+1)}(\xi).$$

We have proved the theorem.

Let us find the expansion of a simple function, an *algebraic polynomial of the nth order*, by Taylor's formula. Suppose that

$$f(x) = C_0 x^n + C_1 x^{n-1} + \ldots + C_{n-1} x + C_n.$$

Then, inasmuch as $f^{(n+1)}(x) \equiv 0$, the remainder $R_{n+1}(x) \equiv 0$ and Taylor's formula (8.33) assumes the form

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f^{(2)}(a)}{2!}(x-a)^2 + \ldots$$

$$\ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n. \qquad (8.44)$$

(In this formula, any point on an infinite line can serve as a.) Thus, Taylor's formula makes it possible to represent any polynomial $f(x)$ *as a polynomial in the powers* $(x-a)$, where a is any real number.

Suppose now that $f(x)$ is an *arbitrary function* satisfying the conditions of Theorem 8.18. Let us try to establish the properties of polynomial (8.35) entering into Taylor's formula for that function. We shall designate this polynomial $\varphi(x, a)$ as before. The symbol $\varphi^{(n)}(x, a)$ will be used to designate the nth derivative $\varphi(x, a)$ with respect to x. Differentiating formula (8.35) with respect to x and putting $x = a$, we obtain the following equalities:

$$\varphi(a, a) = f(a),$$
$$\varphi'(a, a) = f'(a),$$
$$\varphi^{(2)}(a, a) = f^{(2)}(a),$$
$$\ldots\ldots\ldots\ldots$$
$$\varphi^{(n)}(a, a) = f^{(n)}(a).$$

Thus, the polynomial $\varphi(x, a)$ in Taylor's formula for the arbitrary function $f(x)$ possesses the following property: the polynomial itself and its derivatives up to order n inclusive are equal, at point $x = a$, to $f(x)$ and to its derivatives up to order n, respectively.

8.14. VARIOUS FORMS OF THE REMAINDER. MACLAURIN'S FORMULA

8.14.1. The remainder in the form of Lagrange, Cauchy and Peano. Earlier we have established Taylor's formula with the remainder *in a generalized form*. Here we shall establish other possible representations of the remainder. Two of these representations will be obtained as particular cases of the generalized form of the remainder.

First of all, we shall transform somewhat the formula for the remainder (8.34). Since point ξ lies between points a and x, *there can be found a number* θ^* *belonging to the interval* $0 < \theta < 1$ such that $\xi - a = \theta (x - a)$. In this case $\xi = a + \theta (x - a)$, $(x - \xi) = (x - a)(1 - \theta)$. Thus, formula (8.34) can be rewritten in the form

$$R_{n+1}(x) = \frac{(x-a)^{n+1}(1-\theta)^{n-p+1}}{n!p} f^{(n+1)}[a + \theta(x-a)]. \quad (8.45)$$

Let us consider now two important particular cases of formula (8.45): (1) $p = n + 1$, (2) $p = 1$ (recall that in formulas (8.34) and (8.45) any positive number can be taken as p). The first particular case ($p = n + 1$) leads to the remainder *in the form of Lagrange*

$$R_{n+1}(x) = \frac{(x-a)^{n+1}}{(n+1)!} f^{(n+1)}[a + \theta(x-a)]. \quad (8.46)$$

This form of the remainder finds the most frequent use in applications. The remainder in Lagrange's formula resembles the next, in order, term of Taylor's formula in the case where the $(n + 1)$th derivative of the function $f(t)$ is calculated not at point a, but at some intermediate point $\xi = a + \theta (x - a)$ between points a and x. The second particular case mentioned above ($p = 1$) leads to the remainder *in the form of Cauchy*

$$R_{n+1}(x) = \frac{(x-a)^{n+1}(1-\theta)^n}{n!} f^{(n+1)}[a + \theta(x-a)]. \quad (8.47)$$

Since the forms of Lagrange and Cauchy correspond to different values of p, and θ does not depend on p, the values of θ in formulas (8.46) and (8.47) in a general case are *different*. To evaluate some functions, Cauchy's form is preferable to Lagrange's form. Both forms of the remainder (Lagrange's and Cauchy's) can be used for *approximate calculations of the function* $f(x)$ for one or another fixed value of x different from a.

It is natural to replace $f(x)$ by the polynomial $\varphi(x, a)$ and make a numerical evaluation of the error that will be made. But we sometimes encounter problems in which we are interested not in the numerical value of the error made *but in its order with respect to the small value* $(x - a)$. For this purpose, another form of notation for the remainder, the so-called *Peano's*** *form*, is more convenient, and we shall try to establish it in what follows.

Suppose a function $f(x)$ *has derivatives up to the order* $(n - 1)$ *in some neighbourhood of point* a *and a derivative of order* n *at point* a *itself****.

* It should be emphasized that ξ and, consequently, θ depend not only on x and n, but on p as well.

** Guiseppe Peano (1858-1932), an Italian mathematician.

*** Thus we impose on the function $f(x)$ the requirement weaker than the requirements made in Theorem 8.18.

As before, we shall use the symbol $R_{n+1}(x)$ for the difference between the function $f(x)$ and polynomial (8.35) and prove that the equality*

$$R_{n+1}(x) = o[(x-a)^n] \tag{8.48}$$

is valid for $R_{n+1}(x)$.

This last equality is called the remainder *in Peano's form.*

Since with the suppositions we have made, polynomial (8.35) and its derivatives up to order n inclusive coincide, at point $x = a$, respectively, with the function $f(x)$ and with its derivatives, taken at the same point $x = a$, we find that equalities

$$R_{n+1}(a) = 0, \quad R'_{n+1}(a) = 0, \quad \ldots, \quad R^{(n-1)}_{n+1}(a) = 0, \quad R^{(n)}_{n+1}(a) = 0 \tag{8.49}$$

hold true, and it remains for us to prove that equalities (8.49) yield representation (8.48).

For our purpose it is sufficient to prove, with the aid of equality (8.49) that

$$\lim_{x \to a} \frac{R_{n+1}(x)}{(x-a)^n} = 0. \tag{8.50}$$

Since each of the functions $R_{n+1}(x)$ and $(x-a)^n$ is differentiable $(n-1)$ times everywhere in some neighbourhood of point a, equalities (8.49) hold true and any derivative of the function $(x-a)^n$ up to order $(n-1)$ inclusive vanishes *only at point a*, it follows that to evaluate the indeterminate expression on the left-hand side of (8.50) we can successively apply L'Hospital's theorem 8.17 $(n-1)$ times and as a result obtain

$$\lim_{x \to a} \frac{R_{n+1}(x)}{(x-a)^n} = \lim_{x \to a} \frac{R'_{n+1}(x)}{n \cdot (x-a)^{n-1}} = \ldots = \lim_{x \to a} \frac{R^{(n-1)}_{n+1}(x)}{n!\,(x-a)}. \tag{8.51}$$

Taking into account the last but one equality (8.49), we can rewrite (8.51) in the form

$$\lim_{x \to a} \frac{R_{n+1}(x)}{(x-a)^n} = \frac{1}{n!} \lim_{x \to a} \frac{R^{(n-1)}_{n+1}(x) - R^{(n-1)}_{n+1}(a)}{x-a}.$$

Since the derivative $R^{(n)}_{n+1}(a)$ exists and is equal to zero by virtue of the last relation (8.49), the limiting value on the right-hand side of the last equality exists and is equal to zero. The proof of equality (8.50) is complete.

Thus we have completed the derivation of representation (8.48).

---

* The symbol $o[(x-a)^n]$ means an infinitesimal, for $x \to a$, quantity of the order of smallness higher than that of $(x-a)^n$ (see 4.2.3).

In conclusion we shall write out Taylor's formula with the remainder in Peano's form:

$$f(x)=f(a)+\frac{f'(a)}{1!}(x-a)+\ldots+\frac{f^n(a)}{n!}(x-a)^n+o[(x-a)^n]. \tag{8.52}$$

8.14.2. Another form of Taylor's formula. Taylor's formula (8.33) is often represented in another form. Let us put in (8.33) $a=x_0$, $(x-a)=\Delta x$ and take the remainder in Lagrange's form (8.46). In that case $x=x_0+\Delta x$, and we obtain

$$f(x_0+\Delta x)-f(x_0)=\frac{f'(x_0)}{1!}\Delta x+\frac{f^{(2)}(x_0)}{2!}(\Delta x)^2+\ldots$$
$$\ldots+\frac{f^{(n)}(x_0)}{n!}(\Delta x)^n+\frac{f^{(n+1)}(x_0+\theta\Delta x)}{(n+1)!}(\Delta x)^{n+1}. \tag{8.53}$$

(Here θ is some number belonging to the interval $0<\theta<1$.) Taylor's formula (8.53) is a natural generalization of Lagrange's formula (8.11) (see 8.9). Lagrange's formula (8.11) is obtained from formula (8.53) in the particular case $n=0$.

8.14.3. Maclaurin's formula. It is customary to say that *Maclaurin's* formula* is in fact Taylor's formula (8.33) with centre at point $a=0$. Thus, Maclaurin's formula represents a function in the neighbourhood of point $x=0$. Let us write Maclaurin's formula for an arbitrary function $f(x)$ with the remainder in the form of Lagrange, Cauchy, and Peano**:

$$f(x)=f(0)+\frac{f'(0)}{1!}x+\frac{f^{(2)}(0)}{2!}x^2+\ldots+\frac{f^{(n)}(0)}{n!}x^n+R_{n+1}(x), \tag{8.54}$$

where the remainder is

(1) in Lagrange's form

$$R_{n+1}(x)=\frac{x^{n+1}}{(n+1)!}f^{(n+1)}(\theta x)\quad(0<\theta<1), \tag{8.55}$$

(2) in the form of Cauchy***

$$R_{n+1}(x)=\frac{x^{n+1}(1-\theta)^n}{n!}f^{(n+1)}(\theta x)\quad(0<\theta<1), \tag{8.56}$$

---

* Colin Maclaurin (1698-1746), a British mathematician.

** It is assumed in this case that $f(x)$ has an $(n+1)$th derivative in the neighbourhood of point $x=0$, and in the case of Peano's form it has an $(n-1)$th derivative of the remainder in the neighbourhood of point $x=0$ and an nth derivative at point $x=0$ itself.

*** It should be emphasized once again that the values of θ in formulas (8.55) and (8.56) are, in a general case, different.

(3) in Peano's form

$$R_{n+1}(x) = o(x^n). \tag{8.57}$$

(We have used formulas (8.46), (8.47) and (8.48).)

Let us now turn to the discussion of the remainder in Taylor-Maclaurin's formula, to seeking the expansion of the most significant elementary functions by Ma laurin's formula and to considering various applications of this formula.

8.15. EVALUATION OF THE REMAINDER. EXPANSION OF SOME ELEMENTARY FUNCTIONS

8.15.1. Evaluating the remainder for an arbitrary function. Let us evaluate, for an arbitrary function $f(x)$, a remainder in Maclaurin's formula (8.54) taken in Lagrange's form (8.55).

Suppose that the function $f(x)$ under consideration possesses the following property: *there is a real number M such that for all the numbers n and for all the values of the argument x from the neighbourhood of point $x = 0$ being considered, the inequality*

$$|f^{(n)}(x)| \leqslant M \tag{8.58}$$

holds true.

We shall refer to the function possessing the indicated property as the *function the collection of whose all derivatives is bounded in the neighbourhood of point $x = 0$.*

Inequality (8.58) yields

$$|f^{(n)}(\theta x)| \leqslant M, \tag{8.59}$$

and, therefore, it follows from formula (8.55) that

$$|R_{n+1}(x)| = \frac{|x|^{n+1}}{(n+1)!}\,|f^{(n+1)}(\theta x)| \leqslant M\,\frac{|x|^{n+1}}{(n+1)!}.$$

Hence, *we obtain the following universal evaluation of the remainder for the function the collection of whose all derivatives is bounded by the number M in the neighbourhood of point $x = 0$:*

$$|R_{n+1}(x)| \leqslant M\,\frac{|x|^{n+1}}{(n+1)!}. \tag{8.60}$$

Recall that for any fixed x

$$\lim_{n\to\infty} \frac{|x|^{n+1}}{(n+1)!} = 0$$

(see Example 3 in 3.3.3), whence it follows that by choosing a sufficiently large number n we can make the right-hand side of (8.60) as small as possible. This enables us to apply Maclaurin's formula for approximate estimation of a function, possessing the indicated

property, with any preassigned accuracy. We shall cite some examples of functions the collection of whose all derivatives is bounded in the neighbourhood of point $x = 0$.

(1) $f(x) = e^x$, $f^{(n)}(x) = e^x$. The collection of all derivatives of this function is bounded on any closed interval $[-r, r]$ $(r > 0)$ by the number $M = e^r$.

(2) $f(x) = \cos x$ or $f(x) = \sin x$. The collection of all derivatives of each of these functions is bounded everywhere on an infinite line by the number $M = 1$.

8.15.2. Expanding some elementary functions by Maclaurin's formula.

A. $f(x) = e^x$. Since $f^{(n)}(x) = e^x$, $f^{(n)}(0) = 1$ for any n, Maclaurin's formula (8.54) has the form

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + R_{n+1}(x), \tag{8.61}$$

where the remainder in Lagrange's form is equal to

$$R_{n+1}(x) = \frac{x^{n+1}}{(n+1)!} e^{\theta x} \quad (0 < \theta < 1).$$

Since $|e^{\theta x}| < e^r$, we shall obtain, on any closed interval $[-r, +r]$ $(r > 0)$, the following evaluation of the remainder:

$$|R_{n+1}(x)| < \frac{r^{n+1}}{(n+1)!} e^r. \tag{8.62}$$

B. $f(x) = \sin x$. Since $f^{(n)}(x) = \sin\left(x + n\frac{\pi}{2}\right)$, we have

$$f^{(n)}(0) = \sin n\frac{\pi}{2} = \begin{cases} 0 & \text{for an even } n, \\ (-1)^{\frac{n-1}{2}} & \text{for an odd } n, \end{cases}$$

and Maclaurin's formula (8.54) is of the form

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots + (-1)^{\frac{n-1}{2}} \frac{x^n}{n!} + R_{n+2}(x), \tag{8.63}$$

where n is an odd number and the remainder in Lagrange's form is equal to

$$R_{n+2}(x) = \frac{x^{n+2}}{(n+2)!} \sin\left(\theta x + n\frac{\pi}{2} + \pi\right) \quad (0 < \theta < 1).$$

It is evident that the evaluation

$$|R_{n+2}(x)| \leqslant \frac{r^{n+2}}{(n+2)!} \tag{8.64}$$

is valid for the remainder on any closed interval $[-r, +r]$ $(r > 0)$.

C. $f(x)=\cos x$. Since $f^{(n)}(x)=\cos\left(x+n\frac{\pi}{2}\right)$, we have

$$f^{(n)}(0)=\cos n\frac{\pi}{2}=\begin{cases}0 & \text{for an odd } n,\\ (-1)^{\frac{n}{2}} & \text{for an even } n,\end{cases}$$

and Maclaurin's formula (8.54) is of the form

$$\cos x=1-\frac{x^2}{2!}+\frac{x^4}{4!}-\frac{x^6}{6!}+\ldots+(-1)^{\frac{n}{2}}\frac{x^n}{n!}+R_{n+2}(x), \quad (8.65)$$

where n is an even number and the remainder in Lagrange's form is equal to

$$R_{n+2}(x)=\frac{x^{n+2}}{(n+2)!}\cos\left(\theta x+n\frac{\pi}{2}+\pi\right) \quad (0<\theta<1).$$

On any closed interval $[-r, +r]$ $(r>0)$ we obtain evaluation (8.64) for the remainder.

D. $f(x)=\ln(1+x)$. Since

$$f^{(n)}(x)=(-1)^{n-1}\frac{(n-1)!}{(1+x)^n},\ f(0)=0,\ f^{(n)}(0)=(-1)^{n-1}(n-1)!,$$

Maclaurin's formula (8.54) is of the form

$$\ln(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\ldots+(-1)^{n-1}\frac{x^n}{n}+R_{n+1}(x). \quad (8.66)$$

This time we shall write out and evaluate the remainder both *in Lagrange's form and in the form of Cauchy*:

$$R_{n+1}(x)=\frac{(-1)^n x^{n+1}}{(n+1)(1+\theta x)^{n+1}} \text{ (in Lagrange's form)}, \quad (8.67)$$

$$R_{n+1}(x)=(-1)^n x^{n+1}\frac{(1-\theta)^n}{(1+\theta x)^{n+1}}\ * \text{ (in Cauchy's form)}. \quad (8.68)$$

To evaluate the function in $\ln(1+x)$ for the values of x belonging to the closed interval $0\leqslant x\leqslant 1$, it is more convenient to proceed from the remainder in Lagrange's form (8.67). Passing to absolute values in formula (8.67), we obtain for all x's from the interval $0\leqslant x<1$

$$|R_{n+1}(x)|<\frac{1}{n+1}. \quad (8.69)$$

It is evident from estimation (8.69) that for all x's, belonging to the interval $0\leqslant x\leqslant 1$, $R_{n+1}(x)\to 0$ for $n\to\infty$.

* Again pay attention to the fact that in formulas (8.67) and (8.68) the values of θ are, generally speaking, different.

Let us estimate now the function $\ln(1+x)$ for *negative* values of x from the interval $-r \leqslant x \leqslant 0$, where $0 < r < 1$. To make the estimation, we shall proceed from the remainder in Cauchy's form (8.68).

We shall rewrite the remainder in the form

$$R_{n+1}(x) = (-1)^n \left(\frac{1-\theta}{1+\theta x}\right)^n \frac{x^{n+1}}{1+\theta x}. \tag{8.70}$$

Bearing in mind that for the values of x being considered $\frac{1-\theta}{1+\theta x} < 1$ and passing to absolute values in formula (8.70), we shall have

$$|R_{n+1}(x)| < \frac{r^{n+1}}{1-r}. \tag{8.71}$$

Since $r < 1$, evaluation (8.71) enables us to assert that $\lim\limits_{n\to\infty} R_{n+1} = 0$.

E. $f(x) = (1+x)^\alpha$, where α is a real number. Since

$$f^{(n)}(x) = \alpha(\alpha-1)\ \ldots\ (\alpha-n+1)(1+x)^{\alpha-n},$$

$$f^{(n)}(0) = \alpha(\alpha-1)\ \ldots\ (\alpha-n+1),$$

Maclaurin's formula (8.54) has the form

$$(1+x)^\alpha = 1 + \frac{\alpha}{1!}x + \frac{\alpha(\alpha-1)}{2!}x^2 + \ldots$$

$$\ldots + \frac{\alpha(\alpha-1)\ \ldots\ (\alpha-n+1)}{n!}x^n + R_{n+1}(x), \tag{8.72}$$

where the remainder in Lagrange's form is

$$R_{n+1}(x) = \frac{\alpha(\alpha-1)\ \ldots\ (\alpha-n)}{(n+1)!}(1+\theta x)^{\alpha-(n+1)}x^{n+1} \quad (0<\theta<1). \tag{8.73}$$

In a particular case, where $\alpha = n$ is an integer, $R_{n+1}(x) = 0$ and we obtain Newton's binomial formula familiar from the course of elementary mathematics:

$$(1+x)^n = 1 + \frac{n}{1!}x + \frac{n(n-1)}{2!}x^2 + \ldots + x^n. \tag{8.74}$$

If we have to factorize the binomial $(a+x)^n$ instead of the binomial $(1+x)^n$, we can factor a^n out and make use of formula (8.74). Then we shall obtain

$$(a+x)^n = a^n\left(1+\frac{x}{a}\right)^n = a^n\left[1 + \frac{n}{1!}\left(\frac{x}{a}\right) + \right.$$

$$\left. + \frac{n(n-1)}{2!}\left(\frac{x}{a}\right)^2 + \ldots + \left(\frac{x}{a}\right)^n\right].$$

Thus, the general case of Newton's binomial is a particular case of Maclaurin's formula.

F. $f(x) = \arctan x$. We can verify that

$$f^{(n)}(0) = \begin{cases} 0 & \text{for an even } n, \\ (-1)^{\frac{n-1}{2}} (n-1)! & \text{for an odd } n. \end{cases}$$

Thus, Maclaurin's formula (8.54) with the remainder in Peano's form (8.57) has the form

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \ldots + (-1)^{\frac{n-1}{2}} \frac{x^n}{n} + o(x^n).$$

(Here n is an odd number.)

8.16. EXAMPLES OF APPLICATIONS OF MACLAURIN'S FORMULA

8.16.1. Algorithm for calculating the number e. In 3.3.4 we introduced the number e as the limit $\lim\limits_{n\to\infty}\left(1+\frac{1}{n}\right)^n$ and obtained for e a rough estimate* $2 \leqslant e \leqslant 3$. Now we will show how to calculate the number e with any required degree of accuracy. We shall make use of Maclaurin's formula (8.61) and of the evaluation of the remainder (8.62), putting in these formulas $x = r = 1$. We obtain

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \ldots + \frac{1}{n!} + R_{n+1}(1), \tag{8.75}$$

where

$$|R_{n+1}(1)| < \frac{e}{(n+1)!} \leqslant \frac{3}{(n+1)!}. \tag{8.76}$$

Taking a sufficiently large n in formulas (8.75) and (8.76), we shall be able to estimate, with the aid of these formulas, the number e with any required degree of accuracy.

8.16.2. Realization of algorithm for calculating the number e by means of a computer. The algorithm of computing the number e indicated in the previous subsection can be easily realized with the aid of an electronic computer.

---

* See formula (3.7) in Chap. 3.

We present here the result of computing the number e by formula (8.75) for $n = 400$ on the computer BESM-6. The computation was carried out with 600 decimal places. Bearing in mind possible errors of approximation, we have discarded the last 10 decimal places and give here the result of the computation with 590 decimal places*.

2, 718281 828459 045235 360287 471352 662497 757247 093699
959574 966967 627724 076630 353547 594571 382178 525166 427427
466391 932003 059921 817413 596629 043572 900334 295260 595630
738132 328627 943490 763233 829880 753195 251019 011573 834187
930702 154089 149934 884167 509244 761460 668082 264800 168477
411853 742345 442437 107539 077744 992069 551702 761838 606261
331384 583000 752044 933826 560297 606737 113200 709328 709127
443747 047230 696977 209310 141692 836819 025515 108657 463772
111252 389784 425056 953696 770785 449969 967946 864454 905987
931636 889230 098793 127736 178215 424999 229576 351482 208269
895193 668033 182528 869398 496465 105820 939239 829488 793320
36...

Make note of the fact that the computation took about a minute of computer time.

---

* For the readers acquainted with the algorithmic language ALGOL we give a computation program written in this language:

ALGOL system—BESM-6, Variant 10-12-69

```
begin integer i, c, p, n, m; integer array a, b, e [0 : 601];
   m: = 400; marg (39, 50, 39, 10, 0, 0);
   e [0]: = 1; b [0] : = 1;
   for i: = 1 step 1 until 601 do
     a [i]: = b [i]: = e [i] : = 0;
   for n: = 1 step 1 until m do
   begin for i: = 0 step 1 until 600 do
     a [i]: = b [i]; c: = a [0];
     for i = 0 step 1 until 600 do
     begin b [i]: = c ÷ n;
       c: = (c — n × b [i]) × 10 + a [i + 1] end p: = 0
     for i: = 600 step — 1 until 0 do
   begin c: = e [i] + b [i] + p;
     p: = 0
     if c < 10 then e [i]: = c else
     begin e [i]: = c — 10; p: = 1 end
  end
end
for n: — 1 step 1 until 6 do
begin output ('10', 'zd.', e [0]);
   for i: = 1 step 1 until 590 do
   output ('zd', e [i])
  end
end
```

8.16.3. Application of Maclaurin's formula for asymptotic* estimations of elementary functions and for computing limits. Maclaurin's formula is a powerful means for obtaining asymptotic estimations of elementary functions and for calculating limits.

In Chapter 4 we established the following asymptotic formulas for elementary functions:

$$\left.\begin{aligned} &\sin x = x + o(x), \\ &\sqrt[n]{1+x} = 1 + \frac{x}{n} + o(x), \\ &\ln(1+x) = x + o(x), \\ &e^x = 1 + x + o(x), \\ &\cos x = 1 - \frac{x^2}{2} + o(x^2). \end{aligned}\right\} \tag{8.77}$$

Formulas (8.77) represent elementary functions for small values of $|x|$. The first four formulas (8.77) estimate the respective elementary functions with an accuracy to within the members of the *first order of smallness* relative to a small value x while the last formula (8.77) estimates them with an accuracy to within the members of the *second order of smallness* relative to x.

Evaluations (8.77) are sufficient to calculate the simplest limits. However, they are not sufficient to calculate more complicated limits in which a decisive role is played by the members of a higher order of smallness than that of the small quantity x. Thus, for instance, it is impossible to obtain the limiting value

$$\lim_{x \to 0} \frac{\sin x - x}{x^3}, \tag{8.78}$$

from formulas (8.77) because we can infer from the form of the denominator that the decisive role is played here by the members of the *third order of smallness* with respect to x.

Consequently, to calculate fine limits it is necessary to obtain more precise asymptotic evaluations for the functions on the left-hand sides of formulas (8.77).

Such evaluations follow immediately from Maclaurin's formula (8.54) when the remainder in this formula is taken in the form of Peano (8.57). Writing Maclaurin's formulas (8.63), (8.72), (8.66), (8.61) and (8.65) and taking in each formula the remainder in Peano's

* The formula or the estimation characterizing the behaviour of $f(x)$ for $x \to a$ (here for $x \to 0$) is called *asymptotic*.

form, we receive the following asymptotic estimations:

$$\left.\begin{aligned}
&\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots + (-1)^{\frac{n-1}{2}} \frac{x^n}{n!} + o(x^{n+1}),\\
&(1+x)^\alpha = 1 + \frac{\alpha}{1!}x + \frac{\alpha(\alpha-1)}{2!}x^2 + \ldots\\
&\ldots + \frac{\alpha(\alpha-1)\ldots(\alpha-n+1)}{n!}x^n + o(x^n),\\
&\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots + (-1)^{n-1}\frac{x^n}{n} + o(x^n),\\
&e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots + \frac{x^n}{n!} + o(x^n),\\
&\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \ldots + (-1)^{\frac{n}{2}}\frac{x^n}{n!} + o(x^{n+1}).
\end{aligned}\right\} \quad (8.79)$$

(Here, in the first formula (8.79), n is an arbitrary *odd* number and in the last formula (8.79) n is any *even* number.) Formulas (8.79) estimate the respective elementary functions with an accuracy to within the members of *any order n* with respect to the small quantity x. These formulas are an efficient means of computing a number of fine limiting values.

Here are some examples of application of asymptotic formulas (8.79).

1°. Let us consider the limiting value (8.78) discussed above. Applying the first formula (8.79) (taken with $n = 3$), we shall have

$$\lim_{x\to 0}\frac{\sin x - x}{x^3} = \lim_{x\to 0}\frac{x - \frac{x^3}{3!} + o(x^4) - x}{x^3} =$$
$$= \lim_{x\to 0}\left[-\frac{1}{3!} + o(x)\right] = -\frac{1}{3!}.$$

$$2°.\ I = \lim_{x\to 0}\frac{e^{-\frac{x^2}{2}} - \cos x}{x^3 \sin x}.$$

Proceeding from the form of the denominator, we can infer that the 4th order members with respect to x must play a decisive role (because $\sin x = x + o(x)$). Making use of formulas (8.79), we can write

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + o(x^5), \quad (8.80)$$

$$\sin x = x + o(x), \quad (8.81)$$

$$e^z = 1 + z + \frac{z^2}{2} + o(z^2).$$

Consequently, for $z = -\frac{x^2}{2}$ we obtain

$$e^{-\frac{x^2}{2}} = 1 - \frac{x^2}{2} + \frac{x^4}{8} + o(x^4). \tag{8.82}$$

By virtue of formulas (8.80), (8.81) and (8.82), the required limiting value can be rewritten in the form

$$I = \lim_{x \to 0} \frac{1 - \frac{x^2}{2} + \frac{x^4}{8} + o(x^4) - 1 + \frac{x^2}{2} - \frac{x^4}{24}}{x^4 + o(x^2)} =$$

$$= \lim_{x \to 0} \frac{\frac{1}{8} - \frac{1}{24} + \alpha(x)}{1 - \alpha(x)} = \frac{1}{8} - \frac{1}{24} = \frac{1}{12}.$$

(The symbol $\alpha(x)$ denotes here the quantity $\frac{o(x^4)}{x^4}$, which is infinitesimal for $x \to 0$.)

$$3^\circ.\ I = \lim_{x \to 0} \left(\cos x + \frac{x^2}{2}\right)^{\frac{1}{x(\sin x - x)}}.$$

We denote by y the quantity* $y = \left(\cos x + \frac{x^2}{2}\right)^{\frac{1}{x(\sin x - x)}}$. Then, $I = \lim\limits_{x \to 0} y$. Taking logarithm of the expression for y, we have

$$\ln y = \frac{1}{x(\sin x - x)} \ln\left(\cos x + \frac{x^2}{2}\right).$$

Let us calculate

$$\lim_{x \to 0} \ln y = \lim_{x \to 0} \frac{\ln\left(\cos x + \frac{x^2}{2}\right)}{x(\sin x - x)}.$$

Inasmuch as $\cos x = 1 - \frac{x^2}{2} + \frac{x^4}{24} + o(x^5)$ and $\sin x = x - \frac{x^3}{6} + o(x^4)$, we obtain

$$\lim_{x \to 0} \ln y = \lim_{x \to 0} \frac{\ln\left(1 + \frac{x^4}{24} + o(x^5)\right)}{-\frac{x^4}{6} + o(x^5)}.$$

---

* For small x the expression $\left(\cos x + \frac{x^2}{2}\right)$ is obviously *positive*.

Now we shall take into account that $\ln(1+z) = z + o(z)$. From this formula it follows that

$$\ln\left(1+\frac{x^4}{24}+o(x^5)\right)=\frac{x^4}{24}+o(x^4).$$

Thus we have

$$\lim_{x\to 0}\ln y=\lim_{x\to 0}\frac{\frac{x^4}{24}+o(x^4)}{-\frac{x^4}{6}+o(x^5)}=\lim_{x\to 0}\frac{\frac{1}{24}+\frac{o(x^4)}{x^4}}{-\frac{1}{6}+o(x)}=-\frac{1}{4}.$$

Hence

$$I=\lim_{x\to 0} y=e^{-\frac{1}{4}}.$$

SUPPLEMENT TO CHAPTER 8

CALCULATION OF ELEMENTARY FUNCTIONS

In this Supplement we shall study the problem of computing the values of basic elementary functions.

Two kinds of algorithms are used to calculate the values of the indicated functions, the first of which is based on the expansion of the function being calculated by Taylor's formula, and the second on its expansion into a chain or continuous fraction. The first algorithm makes it possible to compile a single program for computing the values of the logarithmic function and the inverse trigonometric functions. The second algorithm is the basis of the universal program for computing the rest basic elementary functions.

In what follows we shall substantiate the indicated algorithms and estimate the number of iterations necessary to ensure the required degree of accuracy.

1. Calculation of the logarithmic function and the inverse trigonometric functions. These functions can be computed with the aid of Taylor's formula. We shall pay special attention to computations of the logarithm and the arctangent. The computations of arc cot x, arc sin x and arc cos x can be easily reduced to the computations of the arc tan x with the aid of the following well-known formulas:

$$\operatorname{arc\,cot} x=\frac{\pi}{2}-\operatorname{arc\,tan} x,\quad \operatorname{arc\,sin} x=\operatorname{arc\,tan}\frac{x}{\sqrt{1-x^2}},$$

$$\operatorname{arc\,cos} x=\operatorname{arc\,cot}\frac{x}{\sqrt{1-x^2}}.$$

1.1. Computation of $\ln a$. Let us represent the number $a>0$ in the following form:

$$a = 2^p M, \tag{8.83}$$

where p is an integer and M satisfies the conditions

$$\frac{1}{2}\leqslant M<1. \tag{8.84}$$

Note that the representation of a in form (8.83) is unique. Making use of formula (8.83), we shall obtain the following expression for $\ln a$:

$$\ln a = p \ln 2 + \ln M. \tag{8.85}$$

Putting

$$M = \frac{1}{\sqrt{2}} \frac{1+x}{1-x} \tag{8.86}$$

and substituting this expression for M into (8.85), we transform formula (8.85) for $\ln a$ as follows:

$$\ln a = \left(p - \frac{1}{2}\right) \ln 2 + \ln \frac{1+x}{1-x}. \tag{8.87}$$

Let us expand the function $\ln \frac{1+x}{1-x}$ by Maclaurin's formula. It is easy to verify that this expansion, with the remainder in the form of Lagrange, looks as follows:

$$\ln \frac{1+x}{1-x} = 2x + \frac{2x^3}{3} + \frac{2x^5}{5} + \ldots + \frac{2x^{2n+1}}{2n+1} + R_{2n+2}(x), \tag{8.88}$$

where

$$R_{2n+2}(x) = -\frac{x^{2n+1}}{2n+2}\left[\frac{1}{(1+\theta x)^{2n+2}} - \frac{1}{(1-\theta x)^{2n+2}}\right], \tag{8.89}$$

and the number θ is contained strictly between zero and unity.

The following formula is used for approximate calculation of $\ln a$:

$$\ln a \approx \left(p - \frac{1}{2}\right) \ln 2 + 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \ldots + \frac{x^{2n+1}}{2n+1}\right). \tag{8.90}$$

This formula can be obtained from (8.87) by replacing $\ln \frac{1+x}{1-x}$ by a part of Maclaurin's formula (8.88) for this function without the remainder $R_{2n+2}(x)$. Note that the number x in the approximate formula (8.90) for $\ln a$ can be determined from formula (8.86) with due account of the restrictions (8.84) imposed on M.

Let us now estimate the error of formula (8.90). Since the approximate value of $\ln a$ computed by formula (8.90) differs from the precise value computed by formula (8.87) by the value of the remainder $R_{2n+2}(x)$, it suffices to estimate this remainder in order to determine the error.

First of all we shall elucidate the range of variation of x. We obtain from formula (8.86)

$$x = \frac{M\sqrt{2}-1}{M\sqrt{2}+1}. \tag{8.91}$$

It follows from (8.91) that for the values of M satisfying inequalities (8.84), the absolute value of x complies with the condition*

$$|x| < 0.172. \tag{8.92}$$

It should be borne in mind that the structure of the remainder $R_{2n+2}(x)$ is such that the negative and positive values of x can be estimated in the same way (it is evident from formula (8.89) that the replacement of x by $-x$ does

* Since x is a function of M, the problem reduces to finding the maximum value of the modulus of function (8.91) on the closed interval $\left[\frac{1}{2}, 1\right]$.

not change the structure of $R_{2n+2}(x)$). Therefore, it is sufficient to evaluate $R_{2n+2}(x)$ for $x \geqslant 0$. Taking this and inequality (8.92) into account, we obtain, replacing the quantity x by the number 0.172, the quantity $\frac{1}{1+\theta x}$ by unity and the quantity $\frac{1}{1-\theta x}$ by the number $\frac{1}{1-0.172}$ on the right-hand side of (8.89), the following evaluation:

$$|R_{2n+2}(x)| \leqslant \frac{(0.172)^{2n+2}}{2n-2}\left[1+\frac{1}{(1-0.172)^{2n+2}}\right].$$

In the last formula we shall put $(0.172)^{2n+2}$ inside the brackets. Since $\frac{0.172}{1-0.172} < 0.208$, the resulting evaluation of $R_{2n+2}(x)$ will be

$$|R_{2n+2}(x)| \leqslant \frac{(0.172)^{2n+2}+(0.208)^{2n+2}}{2n+2}. \tag{8.93}$$

When calculating ln a on a computer*, formula (8.90) is usually taken for $n = 6$. As can be seen from (8.93), the accuracy of calculations for this case is estimated by the number $\frac{(0.172)^{14}+(0.208)^{14}}{14}$, which does not exceed $1.625 \cdot 10^{-10}$.

1.2. Calculation of arc tan α. We may, evidently, restrict ourselves to the case of positive values of the argument, for, putting $|\alpha| = x$, we find

$$\text{arc tan } \alpha = \operatorname{sgn} \alpha \cdot \text{arc tan } x.$$

We want to indicate standard transformations which reduce the calculations of arc tan x for the values of the argument x not smaller than $\frac{1}{8}$ to the calculation of arc tan for the values of the argument smaller than 1/8.

First assume that $x \geqslant 1$. We shall put $y = \text{arc tan } x$, i.e. $x = \tan y$ and $x_1 = \tan(y - \text{arc tan } 1)$. From the last formula we obtain $x_1 = \frac{\tan y - 1}{\tan y + 1} = \frac{x-1}{x+1} < 1$. Since arc tan x = arc tan 1 + arc tan $x_1 = \frac{\pi}{4}$ + arc tan x_1, the calculation of arc tan x for the values $x \geqslant 1$ reduces to the calculation of arc tan x_1 for $0 < x_1 < 1$.

Let us now turn to the case where the argument satisfies the inequalities $\frac{1}{8} \leqslant x < 1$.

Suppose $k_1 = 1$, $k_2 = \frac{1}{2}$, $k_3 = \frac{1}{4}$, $k_4 = \frac{1}{8}$. Evidently, the inequalities

$$k_i \leqslant x \leqslant 2k_i \tag{8.94}$$

hold true for some $i = 1, 2, 3, 4$.

Let us put $y = \text{arc tan } x$, i.e. $x = \tan y$ and $x_i = \tan(y - \text{arc tan } k_i)$. We obtain from this formula

$$x_i = \frac{\tan y - k_i}{1 + k_i \tan y} = \frac{x - k_i}{1 + k_i x}.$$

Since $x > 0$, we have $1 + k_i x > 1$. Besides, in accordance with the right-hand inequality (8.94), we have $x - k_i < 2k_i - k_i = k_i$. Therefore, from the last expression for x_i we obtain an inequality $x_i < k_i$. Since arc tan x = arc tan k_i + + arc tan x_i, the calculation of arc tan x for the values of x satisfying inequalities (8.94) reduces to the calculation of arc tan x_i for $0 < x_i < k_i$.

* This is the right way to calculate ln a on the computer BESM-6.

Repeating the transformations of the argument x described above at most four times, we shall reduce the computation of arc tan x for the values of x from the half-interval $\frac{1}{8} \leqslant x < 1$ to the computation of the arctangent for the values of the argument smaller than $\frac{1}{8}$.

To compute arc tan x for $x < \frac{1}{8}$, we make use of Maclaurin's formula

$$\text{arc tan } x = x - \frac{x^3}{3} + \frac{x^5}{5} - \ldots + (-1)^n \frac{x^{2n+1}}{2n+1} + R_{2n+2}(x).$$

To compute arc tan x, the last formula is usually taken for $n = 6$ and the remainder is neglected*. The program for the computations of the logarithm and of the arctangent is the same. To apply this program for computing the arctangent, the only additional thing to do is to change signs of the adjacent terms $\frac{x^{2n+1}}{2n+1}$.

2. Computing the trigonometric functions, the exponential function and the hyperbolic functions. The computations of these functions are based on the application of **continued fractions**. The properties of these fractions needed for our aim are given below, in Subsection 1.

The calculation of all these functions is connected with a certain continued fraction which can be obtained upon the expansion of the function tanh x. Therefore, we shall study in detail the calculation of the values of the function tanh x and then show how to calculate the rest of the functions.

2.1. Some information on continued fractions. A terminating continued fraction $\frac{P_n}{Q_n}$ is the expansion of the form

$$\frac{P_n}{Q_n} = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \ddots + \cfrac{a_n}{b_n}}}}. \tag{8.95}$$

The quantities $a_1, a_2, \ldots, a_n$ are usually called partial numerators and $b_0, b_1, \ldots, b_n$ partial denominators.

The continued fractions

$$\frac{P_0}{Q_0} = \frac{b_0}{1}, \quad \frac{P_1}{Q_1} = b_0 + \frac{a_1}{b_1}, \quad \frac{P_2}{Q_2} = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2}}, \ldots \tag{8.96}$$

are called convergents of the fraction $\frac{P_n}{Q_n}$.

If we put $P_{-1} = 1$, $Q_{-1} = 0$, we shall obtain from (8.96) for the convergents $\frac{P_k}{Q_k}$ $(k = 1, 2, \ldots, n)$ the following formulas connecting P_k with P_{k-1} and P_{k-2}, and Q_k with Q_{k-1} and Q_{k-2}:

$$\left.\begin{aligned} P_k &= b_k P_{k-1} + a_k P_{k-2}, \\ Q_k &= b_k Q_{k-1} + a_k Q_{k-2}. \end{aligned}\right\} \tag{8.97}$$

* This is the way to carry out computations on BESM-6.

We shall need a special formula for the fraction $\frac{P_n}{Q_n}$ specified by relation (8.95). To establish this formula, we shall compare two convergents $\frac{P_k}{Q_k}$ and $\frac{P_{k-1}}{Q_{k-1}}$. The difference between these fractions is, evidently, equal to

$$\frac{P_k}{Q_k} - \frac{P_{k-1}}{Q_{k-1}} = \frac{P_kQ_{k-1} - Q_kP_{k-1}}{Q_{k-1}Q_k}. \tag{8.98}$$

By virtue of (8.97), the numerator of the right-hand side of (8.98) can be written in the form

$$P_kQ_{k-1} - Q_kP_{k-1} = (b_kP_{k-1} + a_kP_{k-2})\,Q_{k-1} - (b_kQ_{k-1} + a_kQ_{k-2})\,P_{k-1} = \\ = -a_k\,[P_{k-1}Q_{k-2} - Q_{k-1}P_{k-2}]. \tag{8.99}$$

Applying relation (8.99) consecutively for the values of k, $(k-1)$, $(k-2)$, ..., 1 and taking into account that $P_{-1} = 1$, $Q_{-1} = 0$, $Q_0 = 1$, we shall impart to fraction (8.98) the following form

$$\frac{P_k}{Q_k} - \frac{P_{k-1}}{Q_{k-1}} = (-1)^{k+1}\,a_ka_{k-1}\ldots a_1\,\frac{1}{Q_{k-1}Q_k}. \tag{8.100}$$

Since

$$\frac{P_n}{Q_n} = \frac{P_0}{Q_0} + \left(\frac{P_1}{Q_1} - \frac{P_0}{Q_1}\right) + \left(\frac{P_2}{Q_2} - \frac{P_1}{Q_1}\right) + \ldots + \left(\frac{P_n}{Q_n} - \frac{P_{n-1}}{Q_{n-1}}\right),$$

we shall obtain, with the aid of (8.100), the required special formula for the fraction $\frac{P_n}{Q_n}$:

$$\frac{P_n}{Q_n} = b_0 + \frac{a_1}{Q_0Q_1} - \frac{a_1a_2}{Q_1Q_2} + \ldots + (-1)^{n+1}\,\frac{a_1a_2\ldots a_n}{Q_{n-1}Q_n}. \tag{8.101}$$

2.2. Decomposition of the function tanh x into a continued fraction. The method of decomposition of the function tanh x into a continued fraction used here was suggested by Schlömilch* for the decomposition of the function tan x into a continued fraction.

Let us consider a function $y = \cosh\sqrt{x}$ for the values $x > 0$. The following identities are obvious which are obtained by means of successive differentiation of the given function and simple transformations:

$$2\sqrt{x}\,y' = \sinh\sqrt{x}, \quad 2\sqrt{x}\,y'' + \frac{y'}{\sqrt{x}} - \frac{y}{2\sqrt{x}} = 0.$$

From the last relation we can obtain the following identity valid for every $x > 0$:

$$4xy'' + 2y' - y = 0. \tag{8.102}$$

Differentiating successively identity (8.102), we shall have

$$\left.\begin{aligned} &4xy''' + 6y'' - y' = 0, \\ &\ldots\ldots\ldots \\ &4xy^{(n+2)} + (4n+2)\,y^{(n+1)} - y^{(n)} = 0. \end{aligned}\right\} \tag{8.103}$$

* Oskar Schlömilch, Ueber den Kettenbruch für tan x. *Zs. Math. u. Phys.*, 2 (1857), 137-165.

Let us designate the relation $\frac{y^{(n+1)}}{y^{(n)}}$ as u_{n+1}. Then, from relation (8.103) we shall obtain the identity $4xu_{n+2} + 4n + 2 = \frac{1}{u_{n+1}}$, which yields the relation

$$u_{n+1} = \frac{\frac{1}{2}}{2n+1+2xu_{n+2}}. \tag{8.104}$$

Since $u_1 = \frac{y'}{y} = \frac{\tanh \sqrt{x}}{2\sqrt{x}}$, relation (8.104) for $n = 0$ can be written in the following form:

$$\tanh \sqrt{x} = \frac{\sqrt{x}}{1+2xu_2}.$$

Let us replace u_2 on the right-hand side of this formula by its expression obtained by means of (8.104) for $n = 1$. The resulting formula will be

$$\tanh \sqrt{x} = \cfrac{\sqrt{x}}{1+\cfrac{x}{3+2xu_3}}.$$

In the last relation we can replace u_3 by its expression obtained with the aid of (8.104) for $n = 2$. Such operations can be carried out any finite number of times. As a result we shall obtain a decomposition of the function $\tanh \sqrt{x}$ into a continued fraction. Replacing in this expansion $\sqrt{x}$ by x, we shall find the required decomposition of the function $\tanh x$ into a terminating continued fraction. This decomposition has the form

$$\tanh x = \cfrac{x}{1+\cfrac{x^2}{3+\cfrac{x^2}{5+\ddots+\cfrac{x^2}{2n+1+2x^2u_{n+2}}}}}. \tag{8.105}$$

2.3. Calculating the values of the function $\tanh x$. Estimating the error of calculations. When the function $\tanh x$ is calculated on a computer, use is usually made of formula (8.105) in which the member $2x^2u_{n+2}$ is neglected. In this case it is customary to take n equal to 6 ($n = 6$) and to restrict the absolute values of x to the number $\frac{\pi}{4}$.

We shall estimate the error of approximation for any number n.

Let us designate by $\overline{\tanh} x$ the approximate value of the function $\tanh x$ obtained from (8.105) by dropping the member $2x^2u_{n+2}$. To elucidate the accuracy of calculation, we must, evidently, estimate the difference $\tanh x - \overline{\tanh} x$. Make note of the fact that $\tanh x$ and $\overline{\tanh}\ x$ are continued fractions which will be denoted by $\frac{P_{n+1}}{Q_{n+1}}$ and $\frac{\overline{P}_{n+1}}{\overline{Q}_{n+1}}$, respectively.

Let us write out the values of the partial numerators a_i, $\bar{a}_i$ and the partial denominators b_i and $\bar{b}_i$ for these fractions (the bar above the letter will denote the quantities related to the fraction $\frac{\bar{P}_{n+1}}{\bar{Q}_{n+1}}$). We have

$$\left.\begin{aligned} &a_1=\bar{a}_1=x, \quad a_2=\bar{a}_2=x^2, \ldots, a_{n+1}=\bar{a}_{n+1}=x^2, \\ &b_0=\bar{b}_0=0, \quad b_1=\bar{b}_1=1, \ldots, b_n=\bar{b}_n=2n-1, \\ &b_{n+1}=2n+1+2x^2u_{n+2}, \quad b_{n+1}=2n+1. \end{aligned}\right\} \tag{8.106}$$

Since $Q_{-1}=\bar{Q}_{-1}=0$ and $Q_0=\bar{Q}_0=1$ for the fractions $\frac{P_{n+1}}{Q_{n+1}}$ and $\frac{\bar{P}_{n+1}}{\bar{Q}_{n+1}}$, formulas (8.106) and relations (8.97) yield the following equalities:

$$\left.\begin{aligned} &Q_1=\bar{Q}_1, \; Q_2=\bar{Q}_2, \ldots, Q_n=\bar{Q}_n, \\ &Q_{n+1}=(2n+1+2x^2u_{n+2})\,\bar{Q}_n+x^2\bar{Q}_{n-1}, \; \bar{Q}_{n+1}=(2n+1)\,\bar{Q}_n+x^2\bar{Q}_{n-1}. \end{aligned}\right\} \tag{8.107}$$

Let us now represent each of the fractions $\frac{P_{n+1}}{Q_{n+1}}$ and $\frac{\bar{P}_{n+1}}{\bar{Q}_{n+1}}$ as (8.101). It is clear from formulas (8.106) and (8.107) that these representations differ only by the last summands. Therefore, the difference $\frac{P_{n+1}}{Q_{n+1}}-\frac{\bar{P}_{n+1}}{\bar{Q}_{n+1}}$ will be equal to the difference between the last summands in the representations of these fractions by formula (8.101). Since the difference between the fractions under consideration is equal to $\tanh x - \overline{\tanh}\, x$, we can make use of formulas (8.106) and obtain the following formula:

$$\tanh x-\overline{\tanh}\, x=(-1)^{n+2}\,x^{2n+1}\left[\frac{1}{Q_nQ_{n+1}}-\frac{1}{\bar{Q}_n\bar{Q}_{n+1}}\right].$$

With the aid of formulas (8.107) we can reduce this relation to the following form:

$$\tan x-\overline{\tan}\, x=(-1)^{n+1}\frac{x^{2n+1}}{\bar{Q}_n\bar{Q}_{n+1}}\left[\frac{2x^2\bar{Q}_nu_{n+2}}{2x^2\bar{Q}_nu_{n+2}+(2n+1)\,\bar{Q}_n+x^2\bar{Q}_{n-1}}\right]. \tag{8.108}$$

To obtain the required estimation, we shall make use of the following two inequalities which will be proved later on.

The inequality

$$Q_k \geqslant (2k-1)!! \tag{8.109}$$

is valid for any $k \geqslant 1$ *when* $x \geqslant 0$.

For $x>0$ *the quantity* u_{n+2} *is positive*:

$$u_{n+2}>0. \tag{8.110}$$

Let us now pass to estimating the difference $\tanh x - \overline{\tanh}\, x$ for $x>0$. Since for a positive x we have $u_{n+2}>0$ (see (8.110)) and any $\bar{Q}_k>0$ (see (8.109)), the expression in brackets on the right-hand side of equality (8.108) does not exceed unity. Then we obtain, from (8.109), the following inequality:

$$\bar{Q}_n\bar{Q}_{n+1} \geqslant [(2n-1)!!]^2\,(2n+1).$$

Therefore, the estimation

$$|\tanh x - \overline{\tanh}\, x| \leqslant \frac{x^{2n+1}}{[(2n-1)!!]^2\,(2n+1)} \tag{8.111}$$

of the error is valid for any number n at $x > 0$.

We shall consider the estimation of the error for $n = 6$ and for the values of x satisfying the inequalities $0 < x < \frac{\pi}{4}$. At $n = 6$ the number $2n - 1$ is equal to 11 and the number $2n + 1$ is equal to 13. Since $\frac{\pi}{4} < 0.8$, we have $x^{13} < (0.8)^{13} < 5.6 \cdot 10^{-2}$. It is easy to calculate that $11!! = 10\,395$. Therefore, taking into account that $2 \cdot 6 + 1 = 13$, we find from formula (8.111) that the error in the approximate calculation of $\tanh x$ for $n = 6$ does not exceed $4 \cdot 10^{-11}$.

Let us prove now inequalities (8.109) and (8.110).

Proof of inequality (8.109).

We shall first prove the nonnegativity of any $\overline{Q}_k$. Formula (8.106) yields the nonnegativity of $\overline{b}_k$ and $\overline{a}_k$ for any $k \leqslant n$ at $x \geqslant 0$. We have mentioned that $\overline{Q}_{-1} = 0$ and $Q_0 = 1$. This fact and the second formula (8.97) yield the nonnegativity of $\overline{Q}_k$ for any $k \leqslant n$.

From the second formula (8.97) and the nonnegativity of a_k and $\overline{Q}_k$ follows an inequality

$$\overline{Q}_k \geqslant b_k \overline{Q}_{k-1}. \tag{8.112}$$

Since $\overline{Q}_0 = 1$ and $b_k = 2k - 1$ for $1 \leqslant k \leqslant n$, we obtain consecutively from inequality (8.112) $\overline{Q}_1 \geqslant 1$, $\overline{Q}_2 \geqslant 3$, . . ., $\overline{Q}_k \geqslant (2k - 1)!!$. We have established the validity of inequality (8.109).

Proof of inequality (8.110).

It is sufficient to prove the positivity of all the derivatives of the function $y = \cosh \sqrt{x}$ for $x > 0$. In this way we shall obviously prove inequality (8.110) since $u_{n+2} = \frac{y^{(n+2)}}{y^{(n+1)}}$.

Multiplying the last relation (8.103) by $\frac{x^{n-\frac{1}{2}}}{4}$, we can rewrite this relation in the form

$$\left[x^{n+\frac{1}{2}} y^{(n+1)}\right]' = \frac{x^{n-\frac{1}{2}}}{4}\, y^{(n)}. \tag{8.113}$$

We shall now verify that

$$\lim_{x \to 0+0} \left[x^{n+\frac{1}{2}} y^{(n+1)}(x)\right] = 0. \tag{8.114}$$

It is sufficient to ascertain that the quantity

$$x^n y^{(n+1)}(x) \tag{8.115}$$

is bounded for $x \to 0 + 0$. It follows from the relations $y = \cosh \sqrt{x}$ and $y' = \frac{\sinh \sqrt{x}}{2\sqrt{x}}$ that $y(x)$ and $y'(x)$ are bounded at $x \to 0 + 0$. But then it follows from (8.102) that the quantity $xy^2(x)$ is bounded at $x \to 0 + 0$.

By the method of induction we find from the last relation (8.103) that the quantity (8.115) is bounded at $x \to 0 + 0$ for any number n. Thus we have proved relation (8.114).

Let us prove that for any number n the derivative

$$y^{(n)}(x) \tag{8.116}$$

is positive at $x > 0$. It is evident that $y^{(0)}(x) = y(x) = \cosh \sqrt{x}$ is positive at $x > 0$. Assume that for a certain number n the quantity (8.116) is positive at $x > 0$. We can then verify that the quantity $y^{(n+1)}(x)$, too, is positive at $x > 0$. We infer from (8.113) that the derivative on the left-hand side of (8.113) is positive at $x > 0$, i.e. the function $x^{n+\frac{1}{2}} y^{(n+1)}(x)$ increases at $x > 0$. Then it follows from (8.114) that the function is positive at $x > 0$. Thus we have that $y^{(n+1)}(x) > 0$ at $x > 0$ and this is the proof of inequality (8.110).

2.4. Calculation of the hyperbolic sine, hyperbolic cosine and the exponential function. In what follows, the symbol $S_n(t)$ will denote the following continued fraction:

$$S_n(t) = 1 + \cfrac{t}{3 + \cfrac{t}{5 + \ddots + \cfrac{t}{2n+1}}} \tag{8.117}$$

It is customary to calculate this fraction on a computer compiling a program for this purpose. Making use of this program, we can easily compile a program for calculating the hyperbolic tangent, for, as was elucidated in the previous subsection, the approximate value of $\tanh x$ can be found from the formula

$$\tanh x \approx \frac{x}{S_n(x^2)}. \tag{8.118}$$

It was also ascertained in previous subsections that with an increase in n the accuracy of calculations increases and the error tends to zero.

The computation of the functions $\sinh 2x$, $\cosh 2x$, e^{2x} can be reduced to the computation of the hyperbolic tangent with the aid of the formulas

$$\sinh 2x = \frac{2\tanh x}{1 - \tanh^2 x}, \quad \cosh x = \frac{1 + \tanh^2 x}{1 - \tanh^2 x}, \quad e^{2x} = \frac{1 + \tanh x}{1 - \tanh x}.$$

These formulas and relation (8.118) yield the following formulas for the approximate values of the enumerated functions:

$$\sinh 2x \approx \frac{2S_n(x^2) \cdot x}{S_n^2(x^2) - x^2}, \quad \cosh 2x \approx \frac{S_n^2(x^2) + x^2}{S_n^2(x^2) - x^2}, \quad e^{2x} \approx \frac{S_n(x^2) + x}{S_n(x^2) - x}.$$

It is clear that with the aid of these formulas and the program for computing $S_n(t)$ it is easy to compile programs for computing $\sinh 2x$, $\cosh 2x$ and e^{2x}.

2.5. Calculation of trigonometric functions. The decomposition of the functions $\tan x$ into a continued fraction is constructed by analogy with the decomposition of the function $\tanh x$.

Let us consider the function $y = \cos \sqrt{x}$ for positive values of x. The following relations obtained by successive differentiations of this function and simple transformations are obvious:

$$2\sqrt{x}\, y' = -\sin \sqrt{x}; \quad 2\sqrt{x}\, y'' + \frac{y'}{\sqrt{x}} = -\frac{y}{2\sqrt{x}}.$$

From the last relation we obtain an identity

$$4xy'' + 2y' + y = 0.$$

A successive differentiation of this identity yields

$$4xy''' + 6y'' + y' = 0,$$

. . . . . .

$$4xy^{(n+2)} + (4n+2)\,y^{(n+1)} + y^{(n)} = 0.$$

We shall designate the relation $\frac{y^{(n+1)}}{y^{(n)}}$ as u_{n+1}. Then, from the last equality we shall obtain the equality $4xu_{n+2} + 4n + 2 = -\frac{1}{u_{n+1}}$ which yields a relation

$$u_{n+1} = \frac{-\frac{1}{2}}{2n+1+2xu_{n+2}},$$

whence, by a complete analogy with our reasoning concerning the hyperbolic tangent, we shall obtain the following decomposition of the function $\tan x$ into a continued fraction:

$$\tan x = \cfrac{x}{1+\cfrac{-x^2}{3+\cfrac{-x^2}{5+\ddots+\cfrac{-x^2}{2n+1+2xu_{n+2}}}}}.$$

The approximate value of $\tan x$ can be obtained from this formula by dropping the member $2x^2u_{n+2}$. With due account of expression (8.117), this approximate value can be obtained from the formula

$$\tan x \approx \frac{x}{S_n(-x^2)}. \tag{8.119}$$

As in the case of the hyperbolic tangent, we can ascertain that with an increase in n the accuracy of calculations performed by formula (8.119) increases and the error tends to zero.

With the aid of the formulas $\sin 2x = \frac{2\tan x}{1+\tan^2 x}$ and $\cos 2x = \frac{1-\tan^2 x}{1+\tan^2 x}$ known from elementary mathematics and relation (8.119), we obtain the following formulas for computing the approximate values of $\sin 2x$ and $\cos 2x$:

$$\sin 2x \approx \frac{2S_n(-x^2)\cdot x}{S_n^2(-x^2)-x^2}, \quad \cos 2x \approx \frac{S_n^2(-x^2)+x^2}{S_n^2(-x^2)-x^2}.$$

We want to point out in conclusion that the accuracy of calculation of all the functions indicated in the last two subsections for six iterations ($n = 6$ will not be lower than 10^{-11} under the condition that the argument x does not) exceed $\pi/4$ in its absolute value.

CHAPTER 9

GEOMETRIC INVESTIGATION OF THE GRAPH OF A FUNCTION. FINDING THE MAXIMUM AND MINIMUM VALUES OF A FUNCTION

9.1. INTERVALS OF MONOTONICITY OF A FUNCTION. FINDING THE POINTS OF EXTREMUM

9.1.1. Determining the intervals of monotonicity of a function. We have established, in 8.10, a number of conditions ensuring an *increase* (or, respectively, a *decrease*, a *nonincrease* and *nondecrease*) of the function $f(x)$ on a certain interval (a, b). For the sake of convenience, we shall formulate these conditions once again:

1°. For the function $f(x)$ differentiable on the interval (a, b) to be *nonincreasing* (*nondecreasing*) on that interval it is necessary and sufficient for the derivative $f'(x)$ of this function to be nonnegative (nonpositive) everywhere on that interval.

2°. For the differentiable function $f(x)$ to be *increasing* (*decreasing*) on the interval (a, b) it is sufficient for the derivative $f'(x)$ of this function to be positive (negative) everywhere on that interval.

Thus, the investigation of the problem concerning the intervals of monotonicity of a differentiable function $f(x)$ reduces to the *determination of the sign of the first derivative of that function.*

By way of illustration, we shall consider the problem of finding the intervals of monotonicity of the function $f(x) = x^3 - 3x^2 - 4$. Since $f'(x) = 3x^2 - 6x = 3x(x - 2)$, it is obvious that $f'(x)$

is positive for $-\infty < x < 0$,
is negative for $0 < x < 2$,
is positive for $2 < x < +\infty$.

Thus, the function being considered increases on each of the half-lines $(-\infty, 0)$ and $(2, +\infty)$ and decreases on the interval $(0, 2)$. The graph of this function is shown in Fig. 9.1.

9.1.2. Finding the points of a possible extremum. In 8.7.2 we introduced the concept of a *local maximum* (*minimum*) of the function $f(x)$ and established the *necessary condition* for the function $f(x)$ to possess a local maximum (minimum) at a given point. For the sake of convenience, we shall formulate once again the definitions and results established in the indicated section.

Let the function $f(x)$ be defined everywhere in a certain neighbourhood of point c. The function $f(x)$ is said to have a *local maximum* (*minimum*) at point c if there is a neighbourhood of point c within

which the value $f(c)$ is the greatest (the least) among all the values of this function.

A local maximum and a local minimum can be called by a single name *extremum*.

The following theorem establishes a *necessary condition for the extremum of a differentiable function*: *if the function $f(x)$ is differentiable at point c and possesses an extremum at that point, then $f'(c) = 0$.*

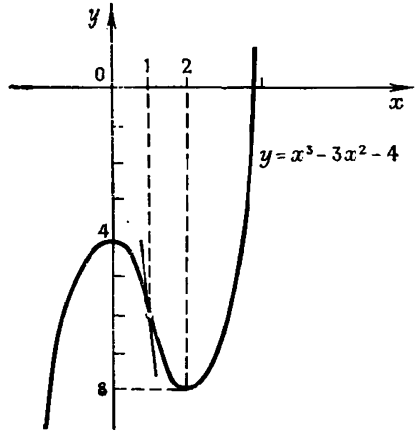


Fig. 9.1

Thus, to find the points of a possible extremum of the differentiable function $f(x)$, it is necessary to find all the roots of the function $f'(x) = 0$ (that is, to find all the zeros of the derivative $f'(x)$). In what follows, we shall refer to the roots of the equation $f'(x) = 0$ as the points of a *possible extremum* of the function $f(x)$*.

Bear in mind, however, that the equality to zero of the first derivative is *only a necessary*** condition for the extremum, we also have to investigate the problem of the existence of an extremum at each point of a possible extremum. To conduct this additional investigation, it is necessary to establish the *sufficient conditions for the existence of an extremum*, and we shall now busy ourselves with this problem.

9.1.3. The first sufficient condition of extremum.

Theorem 9.1. *Suppose point c is a point of a possible extremum of the function $f(x)$ and assume that the function $f(x)$ is differentiable everywhere in a certain neighbourhood of point c. Then, if, within the indicated neighbourhood, the derivative $f'(x)$ is positive (negative) to the left of point c and negative (positive) to the right of point c, the function $f(x)$ has a local maximum (minimum) at point c. Now if the derivative $f'(x)$ is of the same sign to the left and to the right of point c, there is no extremum at point c at all.*

Proof. (1) Suppose first that within the neighbourhood being considered the derivative $f'(x)$ is positive (negative) to the left of c and negative (positive) to the right of c. We are to prove that the value $f(c)$ is the greatest (least) of all the values of $f(x)$ in the indicated neighbourhood. We shall denote by x_0 any value of the argu-

* The roots of the equation $f'(x) = 0$ are sometimes called stationary points.

** You may consider the function $y = x^3$, for instance, to verify that this condition is not sufficient. This function does not have an extremum at point $x = 0$ at which $f'(x) = 0$.

ment from this neighbourhood distinct from c. It is sufficient to prove that

$$f(c) - f(x_0) > 0 \qquad (<0).$$

The function $f(x)$ is differentiable (and, hence, continuous) on the closed interval $[c, x_0]$. Applying Lagrange's Theorem 8.12 to $f(x)$ with respect to the interval $[c, x_0]$, we have

$$f(c) - f(x_0) = f'(\xi)(c - x_0), \tag{9.1}$$

where ξ is a value of the argument lying between c and x_0. Since the derivative $f'(\xi)$ is positive (negative) for $x_0 < c$ and negative (positive) for $x_0 > c$, the right-hand side of (9.1) is positive (negative).

(2) Suppose now that the derivative $f'(x)$ is of the same sign to the left and to the right of c. Designating, as before, by x_0 *any* value of the argument different from c and repeating the argument given above, we shall prove that the right-hand side of (9.1) has *unlike* signs for $x_0 < c$ and $x_0 > c$. This proves that there is no extremum at point c.

A rule yielded by Theorem 9.1 can be briefly formulated as follows: (1) *if, when passing through a given point c of a possible extremum, the derivative $f'(x)$ changes sign from plus to minus (from minus to plus), the function $f(x)$ possesses at point c a local maximum (minimum)*; (2) *now if the derivative $f'(x)$ does not change sign when passing through the given point c of a possible extremum, then there is no extremum at point c.*

Examples. (1) Assuming that a tin has a form of a circular cylinder of radius r and height h, determine at what ratio between r and h the tin with a constant area of the entire surface will have the greatest volume.

Let us designate the area of the entire surface of the tin by S. Then we have

$$2\pi r^2 + 2\pi rh = S = \text{const}. \tag{9.2}$$

We find from this equality that

$$h = \frac{S}{2\pi r} - r.$$

Thus we can express the volume V of the tin as a function of the radius r:

$$V = \pi r^2 h = \frac{S}{2} r - \pi r^3.$$

We have reduced the problem to seeking the maximum of the function

$$V(r) = \frac{S}{2} r - \pi r^3.$$

Equating the derivative $V'(r)=\frac{S}{2}-3\pi r^2$ to zero and taking into account that $r>0$, we find the point of a possible extremum

$$r=\sqrt{\frac{S}{6\pi}}. \tag{9.3}$$

Although it is clear from the meaning of the problem that the only point of a possible extremum is the point of maximum of the function $V(r)$, we can strictly verify this by making use of Theorem 9.1 and noting that the derivative $V'(r)=3\pi\left(\frac{S}{6\pi}-r^2\right)$ is positive for $r<\sqrt{\frac{S}{6\pi}}$ and negative for $r>\sqrt{\frac{S}{6\pi}}$. Let us now determine at what ratio between the radius r and the height h the volume $V(r)$ of the tin is the greatest. For this purpose, we shall divide equality (9.2) by r^2 and make use of relation (9.3) on the right-hand side of the resulting equality. We shall obtain

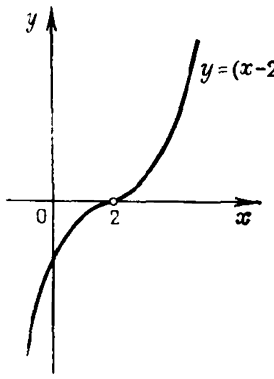


Fig. 9.2

$$\frac{h}{r}=2, \text{ i.e. } h=2r.$$

Thus *the tin with the height equal to the diameter will have the greatest volume**.

(2) Find the extremal points of the function $f(x)=(x-2)^5$. Since $f'(x)=5(x-2)^4$, the only point of a possible extremum is the point $x=2$.

Inasmuch as $f'(x)$ is positive both to the left and to the right of that point, the function $f(x)=(x-2)^5$ has no points of extremum at all (the graph of the function $f(x)=(x-2)^5$ is shown in Fig. 9.2).

9.1.4. The second sufficient condition of extremum. It is sometimes rather difficult to investigate the sign of the first derivative $f'(x)$ to the left and to the right of the point of a possible extremum. For that case, we will indicate another sufficient condition for the existence of an extremum at a given point c of a possible extremum which does not require the investigation of the sign of $f'(x)$ in the neighbourhood of c but, on the other hand, *presupposes the existence at point c of a finite second derivative $f^{(2)}(x)$ different from zero.*

Theorem 9.2. *Suppose the function $f(x)$ possesses a finite second derivative at a given point c of a possible extremum. Then, the function $f(x)$ possesses at point c a maximum if $f^{(2)}(c)<0$ and a minimum if $f^{(2)}(c)>0$.*

* The problem we have solved shows that to economize in the material it is most expedient to make tins with the height equal to the diameter.

Proof. It follows from the condition $f^{(2)}(c) < 0$ (>0) and from Theorem 8.9 that the function $f'(x)$ decreases (increases) at point c. Since $f'(c) = 0$ by the hypothesis, there is a neighbourhood of point c within which $f'(x)$ is positive (negative) to the left of c and negative (positive) to the right of c. But then, in accordance with the previous theorem, $f(x)$ possesses a maximum (minimum) at point c.

Remark. Generally speaking, Theorem 9.2 has a narrower field of application than Theorem 9.1. Thus, Theorem 9.2 does not solve the problem of extremum for the case where the second derivative

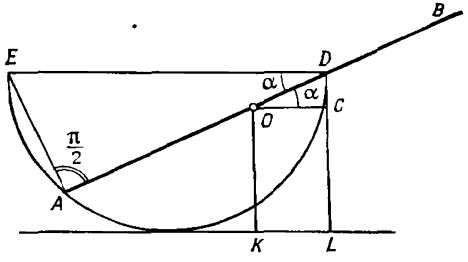


Fig. 9.3

$f^{(2)}(x)$ does not exist at point c and also for the case of $f^{(2)}(c) = 0$. In the latter case, to solve the problem of the existence of an extremum, it is necessary to investigate the behaviour of the higher-order derivatives at point c. This investigation will be done in Sec. 4 of the present chapter.

Examples. (1) A homogeneous bar of length l is descended into a cup having the form of a hemisphere of radius r (Fig. 9.3). Find the position of equilibrium of the bar assuming that $2r < l < 4r$.

The position of equilibrium of the bar corresponds to its minimum potential energy, that is, to the lowest position of its centre of gravity O (since the bar is homogeneous, its centre of gravity coincides with its middle point). Denoting by OK the perpendicular to the plane on which the cup rests, we reduce the problem to searching the position of the bar AB at which the segment OK has the minimum length. We shall first calculate the length of the segment OK as a function of the angle of inclination α of the bar to the plane supporting the cup. Suppose DL is parallel to OK and OC is perpendicular to OK (D being the point at which the bar bears against the edge of the cup).

It follows from the right triangle EAD that $AD = ED \cos \alpha = 2r \cos \alpha$. By the hypothesis, $AO = l/2$. Thus we have

$$OD = AD - AO = 2r \cos \alpha - \frac{l}{2}.$$

On the other hand, $DC = DL - OK = r - OK$. Therefore, it follows from the right triangle ODC that

$$\sin\alpha = \frac{DC}{OD} = \frac{r - OK}{2r\cos\alpha - \frac{l}{2}}.$$

Thus, the length of the segment OK, which will be denoted as $f(\alpha)$, is equal to

$$f(\alpha) = r + \frac{l}{2}\sin\alpha - r\sin 2\alpha.$$

Now we proceed to seeking the value of the angle α which yields the minimum of $f(\alpha)$. (It is clear that we can restrict ourselves to the values of the angle α from the first quadrant.) Since

$$f'(\alpha) = \frac{l}{2}\cos\alpha - 2r\cos 2\alpha = \frac{l}{2}\cos\alpha + 2r - 4r\cos^2\alpha,$$

the points of a possible extremum can be found as solutions of the quadratic equation

$$4r\cos^2\alpha - \frac{l}{2}\cos\alpha - 2r = 0.$$

Since in the first quadrant $\cos\alpha$ is positive, we can use only the positive root of this equation:

$$\cos\alpha_0 = \frac{l + \sqrt{l^2 - 128r^2}}{16r}. \tag{9.4}$$

Though it is clear from the formulation of the problem that the only point of a possible extremum α_0 is the point of minimum of the function $f(\alpha)$, we shall determine this rigorously with the aid of Theorem 9.2. It is sufficient to verify that $f^{(2)}(\alpha_0) > 0$. Inasmuch as

$$f^{(2)}(\alpha) = -\frac{l}{2}\sin\alpha + 4r\sin 2\alpha = 8r\sin\alpha\left(\cos\alpha - \frac{l}{16r}\right),$$

we have, in accordance with (9.4),

$$f^{(2)}(\alpha_0) = 8r\sin\alpha_0\left(\cos\alpha_0 - \frac{l}{16r}\right) = \frac{\sin\alpha_0}{2}\sqrt{l^2 + 128r^2} > 0.$$

Thus we have determined that the position of equilibrium of the bar corresponds to its angle of inclination to the plane supporting the cup specified by formula (9.4).

(2) Find the extremal values of the function $f(x) = x^3 - 3x^2 - 4$. We have dealt with this function in 9.1.1 (see Fig. 9.1). Since

$$f'(x) = 3x^2 - 6x = 3x(x - 2),$$

the function $f(x)$ possesses two points of a possible extremum: $x_1 = 0$ and $x_2 = 2$. Since the sign of $f'(x)$ to the left and to the right

of these points can be easily verified, the question concerning the extremum can be solved with the aid of Theorem 9.1 (the first sufficient condition). But we think it better to apply Theorem 9.2 (the second sufficient condition). We have

$$f^{(2)}(x) = 6x - 6, \quad f^{(2)}(0) = -6 < 0, \quad f^{(2)}(2) = 6 > 0.$$

Thus we see that the function $f(x)$ possesses a maximum at point 0 and a minimum at point 2. The extremal values of this function are equal to

$$f_{\max} = f(0) = -4, \quad f_{\min} = f(2) = -8.$$

9.1.5. Extremum of a function nondifferentiable at a given point. A general scheme of searching for extrema. Up till now we were solving a question of the existence of an extremum of the function $f(x)$ *at a point c at which the function $f(x)$ is differentiable.* Here we shall deal with the problem of the existence of an extremum at point c of a function which is not differentiable at point c but differentiable everywhere in a certain neighbourhood to the right and to the left of point c.

It turns out that Theorem 9.1 can be generalized to the case of such a function. Hence, we can make the following assertion.

Theorem 9.3. *Assume that the function $f(x)$ is differentiable everywhere in a certain neighbourhood of point c, with a possible exception of point c itself, and continuous at point c.*

Then, if in the indicated neighbourhood the derivative $f'(x)$ is positive (negative) to the left of point c and negative (positive) to the right of point c, the function $f(x)$ possesses at point c a local maximum (minimum). Now if the derivative $f'(x)$ has the same sign to the left and to the right of point c, there is no extremum at point c.

The **proof** is completely analogous to that of Theorem 9.1. But this time the applicability of Lagrange's theorem to the function $f(x)$ with respect to the closed interval $[c, x_0]$ is established as follows: by the hypothesis the function $f(x)$ is differentiable (and, hence, continuous) everywhere on the half-open interval $(c, x_0]$ and, besides, is continuous at point c. Consequently, $f(x)$ is continuous everywhere on the interval $[c, x_0]$ and differentiable at all the interior points of that interval.

Examples. (1) Find the points of extremum of the function $f(x) = |x|$. This function is differentiable everywhere on an infinite straight line, except for point $x = 0$, and continuous at point $x = 0$, with the derivative $f'(x)$ being equal to unity for $x > 0$ and to -1 for $x < 0$.

Theorem 9.1 cannot be applied to this function, but according to Theorem 9.3 the function has a minimum at $x = 0$ (Fig. 9.4).

(2) Find the extremal points of the function $y = x^{2/3}$. This function is continuous throughout the infinite line and differentiable

everywhere on that line, except for point $x = 0$. For $x \neq 0$ the derivative is equal to

$$y' = \frac{2}{3} \frac{1}{\sqrt[3]{x}}.$$

In the previous example the derivative had a discontinuity of the 1st kind at point $x = 0$*; this time the derivative has a discontinuity of the 2nd kind ("infinite jump"). We infer from the expression for the derivative that the derivative is negative to the left of point

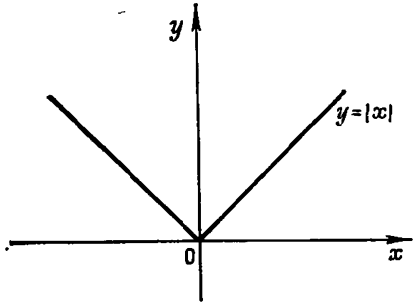


Fig. 9.4

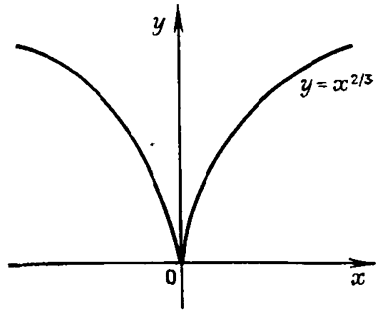


Fig. 9.5

$x = 0$ and positive to the right of that point. Consequently, Theorem 9.3 enables us to assert that the function under consideration has a minimum at point $x = 0$ (the graph of this function is shown in Fig. 9.5).

(3) Find the extremal points of the function

$$y = f(x) = \begin{cases} \dfrac{x}{1+e^{1/x}} & \text{for } x \neq 0, \\ 0 & \text{for } x = 0. \end{cases}$$

It is easy to see that the function is continuous throughout the infinite line. Indeed, the only "doubtful" point is point $x = 0$, but at that point, too, the function is continuous since

$$\lim_{x \to 0+0} y = \lim_{x \to 0-0} y = 0.$$

Besides, it is evident that the function being discussed is differentiable on the entire infinite line, except for point $x = 0$. Everywhere with the exception of that point the derivative is specified by the

* In the sense that although the derivative did not exist at point $x = 0$, it possessed at that point finite right-hand and left-hand limiting values not coinciding with each other.

formula

$$y' = \frac{1 + e^{1/x} + \frac{1}{x} e^{1/x}}{(1+e^{1/x})^2}.$$

It is easy to see that the limit $\lim\limits_{x \to 0} \frac{f(x)-f(0)}{x} = \lim\limits_{x \to 0} \frac{1}{1+e^{1/x}}$ does not exist so that the function $y = f(x)$ is not differentiable at point $x = 0$. Since the derivative of y is positive both to the left and to the right of point $x = 0$, in accordance with Theo-

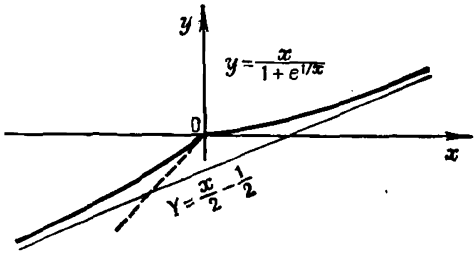


Fig. 9.6

rem 9.3 the function being considered does not have an extremum at point $x = 0$ and, consequently, has no extremum at all. (The graph of the function under consideration is shown in Fig. 9.6.)

We shall give now a general scheme of searching for a local extremum. Let us assume that the function $f(x)$ *is continuous on an interval** (a, b) *and that its derivative* $f'(x)$ *exists and is continuous on that interval everywhere except for a finite number of points.*

Let us suppose, in addition, that the *derivative* $f'(x)$ *vanishes on the interval* (a, b) *only at a finite number of points.* In other words, we postulate that there is only a finite number of points on the interval (a, b) at which the derivative $f'(x)$ either does not exist or turns into zero. We designate these points as $x_1, x_2, \ldots, x_n$ $(a < x_1 < x_2 < \ldots < x_n < b)$. By virtue of our suppositions, the derivative $f'(x)$ *retains a constant sign* on each of the intervals (a, x_1), $(x_1, x_2), \ldots, (x_n, b)$. Consequently, *the question of the existence of an extremum at each of the points* $x_1, x_2, \ldots, x_n$ *can be solved (either in the affirmative or in the negative) with the aid of Theorem 9.3.*

We will not illustrate here the general scheme of searching for the points of a local extremum, the example will be furnished in 9.6.

* Instead of the interval (a, b) we can consider a half-line, an infinite line or some other set.

9.2. THE DIRECTION OF THE CONVEXITY OF THE GRAPH OF A FUNCTION

Suppose that the function $f(x)$ is differentiable at any point belonging to the interval (a, b). Then, as was established in 5.1.4, *there is a tangent to the graph of the function* $y = f(x)$ *passing through any point* $M(x, f(x))$ *of that graph* $(a < x < b)$, *the tangent not being parallel* to the y-axis.*

Definition. *The graph of the function* $y = f(x)$ *is said to be convex downwards (upwards) on the interval* (a, b) *if, within the indicated*

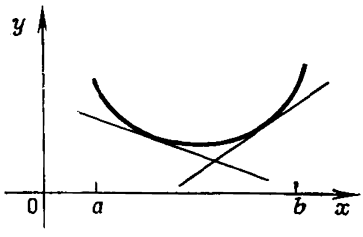


Fig. 9.7

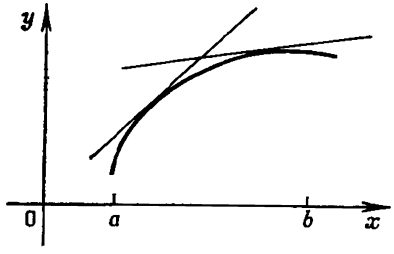


Fig. 9.8

interval, the graph of that function lies not lower (not higher) than any of its tangents.

Remark 1. The expression "the graph lies not lower (not higher) than its tangent" is meaningful since the tangent is not parallel to the y-axis.

Fig. 9.7 shows the graph of a function convex downwards on the interval (a, b) and Fig. 9.8 shows the graph of a function convex upwards.

Theorem 9.4. *If the function* $y = f(x)$ *possesses a finite second derivative on the interval* (a, b) *and if this derivative is nonnegative (nonpositive) everywhere on that interval, then the graph of the function* $y = f(x)$ *is convex downwards (upwards) on the interval* (a, b).

Proof. For the sake of definiteness, we shall consider the case where the second derivative $f^{(2)}(x) \geqslant 0$ everywhere on (a, b). We shall designate as c *any* point belonging to the interval (a, b) (Fig. 9.9). We are to prove that the graph of the function $y = f(x)$ lies not lower than the tangent passing through point $M(c, f(c))$. Let us write an equation of the indicated tangent designating its running ordinate by Y. Since the slope of the indicated tangent is equal to $f'(c)$, its equation is of the form**

$$Y - f(c) = f'(c)(x - c). \tag{9.5}$$

* Since its gradient equal to $f'(x)$ is finite.

** It was pointed out in A. V. Pogorelov's *Analytical Geometry*, Mir Publishers, that the equation of a straight line passing through point $M(a, b)$ and having a slope k is of the form $Y - b = k(x - a)$.

Let us expand the function $f(x)$ in the neighbourhood of point c into Taylor's series, putting in the formula $n = 1$. We obtain

$$y = f(x) = f(c) + \frac{f'(c)}{1!}(x-c) + \frac{f^{(2)}(\xi)}{2!}(x-c)^2, \tag{9.6}$$

where the remainder is taken in Lagrange's form and ξ is contained between c and x. Since by the hypothesis $f(x)$ possesses a second

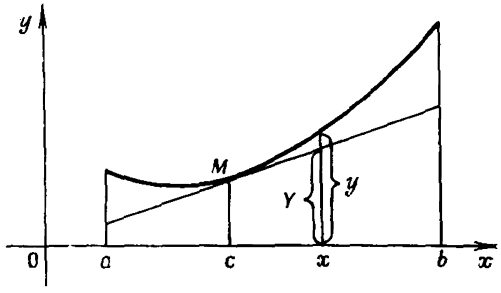


Fig. 9.9

derivative on the interval (a, b), formula (9.6) is valid for *any* x belonging to the interval (a, b) (see 8.13).

Comparing (9.6) and (9.5), we have

$$y - Y = \frac{f^{(2)}(\xi)}{2}(x-c)^2. \tag{9.7}$$

Since by the hypothesis the second derivative $\geqslant 0$ everywhere on (a, b), the right-hand side of (9.7) is *nonnegative*, that is, for all x from (a, b) $y - Y \geqslant 0$ or $y \geqslant Y$.

The last inequality proves that the graph of the function $y = f(x)$ lies not lower than tangent (9.5) everywhere within the interval (a, b).

The proof of the theorem for the case $f^{(2)}(x) \leqslant 0$ is similar.

Remark 2. If $f^2(x) = 0$ everywhere on the interval (a, b), then, as it is easy to verify, $y = f(x)$ is a linear function, i.e. its graph is a straight line. In that case, the direction of convexity can be assumed to be arbitrary.

Theorem 9.5. Assume that the second derivative of the function $y = f(x)$ is continuous and positive (negative) at point c. Then, there exists a neighbourhood of point c throughout which the graph of the function $y = f(x)$ is convex downwards (upwards).*

Proof. According to Theorem 8.4 on the constancy of sign of a continuous function, there is a neighbourhood of point c within which the second derivative $f^{(2)}(x)$ is positive (negative). In accor-

* Recall that the neighbourhood of point c is the interval containing point c.

dance with the preceding theorem, in this neighbourhood the graph of the function $y = f(x)$ is convex downwards (upwards).

Thus, the direction of convexity of the graph of a function is completely defined by the sign of the second derivative of that function.

Example. Determine the direction of convexity of the graph of the function $y = f(x) = x^3 - 3x^2 - 4$. We have already considered this function in 9.1.1 and 9.1.4 (see Fig. 9.1). It follows from the form of the second derivative $f^{(2)}(x) = 6x - 6 = 6(x - 1)$ that the derivative is negative for $x < 1$ and positive for $x > 1$. Thus, the graph of the function $y = x^3 - 3x^2 - 4$ is convex upwards on the interval $(-\infty, 1)$ and downwards on the interval $(1, \infty)$.

9.3. POINT OF INFLECTION OF THE GRAPH OF A FUNCTION

9.3.1. Definition of the point of inflection. A necessary condition for inflection. Assume that a, b and c are three arbitrary numbers connected by the inequalities $a < c < b$. Suppose that the function

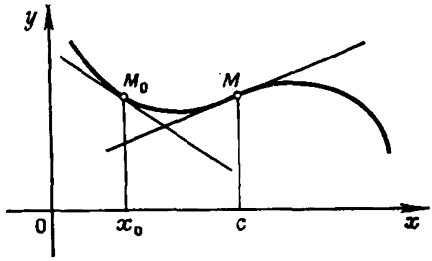


Fig. 9.10

$y = f(x)$ is differentiable on the interval (a, b), that is, there exists a tangent to the graph of that function at all points whose abscissas belong to the interval (a, b). Assume, moreover, that the convexity of the graph of the function $y = f(x)$ has a definite direction on each of the intervals (a, c) and (c, b).

Definition. *Point $M(c, f(c))$ of the graph of the function $y = f(x)$ is called a* **point of inflection** *of that graph if there exists a neighbourhood of point c of the abscissa-axis such that within this neighbourhood the convexity of the graph of the function $y = f(x)$ to the left and to the right of point c has different directions.*

Fig. 9.10 shows the graph of a function possessing an inflection at point $M(c, f(c))$.

Sometimes the definition of a point of inflection of the graph of the function $y = f(x)$ includes an additional requirement that the *indicated graph should lie on different sides of the tangent to that graph, at point $M(c, f(c))$, everywhere within a sufficiently small neighbourhood of point c of the abscissa-axis to the left and to the right of point c.*

We shall show later on that this property follows from the definition we have given under the assumption that the derivative $f'(x)$ is continuous at point c.

Let us prove the following two lemmas.

Lemma 1. *Suppose the function $y = f(x)$ possesses a derivative $f'(x)$ everywhere in the δ-neighbourhood of point c, the derivative being continuous at point c. Then, if the graph of the function $y = f(x)$ has a convexity directed downwards (upwards) on the interval $(c, c + \delta)$, then everywhere within the interval $(c, c + \delta)$ this graph lies not lower (not heigher) than the tangent to the graph constructed at point $M(c, f(c))$.*

Proof. Let us consider the sequence $\{x_n\}$ of the points from the interval $(c, c + \delta)$ converging to point c. Through every point $M_n(x_n, f(x_n))$ of the graph of the function $y = f(x)$ we shall draw a tangent to that graph, i.e. a straight line*

$$Y_n = f(x_n) + f'(x_n)(x - x_n).$$

Since by the hypothesis the graph of the function $y = f(x)$ has a convexity directed downwards (upwards) on the interval $(c, c + \delta)$, we have

$$f(x) - Y_n = f(x) - f(x_n) - f'(x_n)(x - x_n) \geqslant 0 \quad [\leqslant 0]. \qquad (*)$$

for any number n and for any *fixed* point x from the interval $(c, c + \delta)$.

It follows from the condition of continuity of $f'(x)$ (and, the more so, of $f(x)$) at point c and from definition 1 of continuity given in 4.3.1, that there exists a limit

$$\lim_{n\to\infty}(f(x) - Y_n) = \lim_{n\to\infty}\{f(x) - f(x_n) - f'(x_n)(x - x_n)\} =$$
$$= f(x) - f(c) - f'(c)(x - c).$$

By virtue of inequality (*) and Theorem 3.13 from 3.1, it follows from the existence of the last limit that

$$f(x) - f(c) - f'(c)(x - c) \geqslant 0 \quad [\leqslant 0].$$

If we denote by Y the running ordinate of the tangent (9.5) passing through point $M(c, f(c))$, the last inequality can be rewritten in the form $f(x) - Y \geqslant 0$ $[\leqslant 0]$.

Thus, passing to the limit in inequality (*) for $n \to \infty$ and making use of Theorem 3.13 from Chapter 3, we obtain that

$$f(x) - Y \geqslant 0 \quad [\leqslant 0]$$

* We make use of the equation of a straight line passing through a given point $M_n(x_n, f(x_n))$ and having a slope equal to $f'(x_n)$. The running ordinate of that line will be denoted by Y_n.

for any fixed point x belonging to the interval $(c, c + \delta)$, with Y denoting the running ordinate of the tangent drawn through point $M(c, f(c))$. We have proved the lemma.

Remark. The formulation and proof of Lemma 1 are quite analogous for the case where the convexity of the graph of the function has a certain direction not on the interval $(c, c + \delta)$ but on the interval $(c - \delta, c)$.

Lemma 2. *Suppose the function $y = f(x)$ has a derivative $f'(x)$ in some neighbourhood of point c, the derivative being continuous at point c. In that case, if the graph of the function $y = f(x)$ has an inflection at point $M(c, f(c))$, then within a sufficiently small δ-neighbourhood of point c, to the left and to the right of c, this graph lies on different sides of the tangent drawn through point $M(c, f(c))$.*

To prove the lemma, we must choose $\delta > 0$ small enough for the convexity of the graph on the function $y = f(x)$ on each of the intervals $(c - \delta, c)$ and $(c, c + \delta)$ to have a certain direction (which will be different on the intervals $(c - \delta, c)$ and $(c, c + \delta)$). Then, to prove Lemma 2, we only have to apply Lemma 1 to the function $y = f(x)$ on each of the intervals $(c - \delta, c)$ and $(c, c + \delta)$.

Lemma 2 makes it possible to establish a necessary condition of inflection of the graph of a function twice differentiable at a given point.

Theorem 9.6. (a necessary condition of inflection of the graph of a twice differentiable function). *If the function $y = f(x)$ possesses a second derivative at point c and the graph of that function has an inflection at point $M(c, f(c))$, then $f^{(2)}(c) = 0$.*

Proof. Assume, as before, that Y is a running ordinate of the tangent line $Y = f(c) + f'(c)(x - c)$ passing through point $M(c, f(c))$ of the graph.

Let us consider the function

$$F(x) = f(x) - Y = f(x) - f(c) - f'(c)(x - c),$$

equal to the difference between $f(x)$ and the linear function $f(c) + f'(c)(x - c)$.

This function $F(x)$, like the function $f(x)$, possesses a second derivative at point c (and, therefore, possesses a first derivative in some neighbourhood of c, this first derivative being continuous at point c). By virtue of Lemma 2, in a small neighbourhood of point c the graph of the function $y = f(x)$ lies to the left and to the right of c on both sides of the tangent running through point $M(c, f(c))$ and so the function $F(x)$ in the small neighbourhood of point c has *unlike signs* to the left and to the right of c.

Consequently, *the function $F(x)$ cannot have a local extremum at point c.*

Now suppose that $f^{(2)}(c) \neq 0$. Then, since $F'(x) = f'(x) - f'(c)$, $F^{(2)}(x) = f^{(2)}(x)$, the conditions $F'(c) = 0$ and $F^{(2)}(c) \neq$

$\neq 0$ hold true and the function $F(x)$ possesses a local extremum at point c by virtue of Theorem 9.2. The contradiction obtained shows that the supposition $f^{(2)}(c) \neq 0$ is incorrect, i.e. $f^{(2)}(c) = 0$. The proof of the theorem is complete.

The fact that the vanishing of the second derivative is *only a necessary* condition of inflection of the graph of a twice differentiable function follows, for instance, from the consideration of the graph of the function $y = x^4$. For that function, the second derivative $y^{(2)} = 12x^2$ vanishes at point $x = 0$ but its graph has no inflection at point $M(0, 0)$.

In accordance with Theorem 9.6, to find all the points of inflection of the graph of the twice differentiable function $y = f(x)$, it is necessary to consider all the roots of the equation $f^{(2)}(x) = 0$.

Since the equality to zero of the second derivative is only a necessary condition of inflection, we must additionally discuss the problem of the existence of inflection at each point for which $f^{(2)}(x) = 0$. For our purpose we must establish sufficient conditions of inflection and we shall turn now to this question.

9.3.2. The first sufficient condition of inflection.

Theorem 9.7. *Assume that the function $y=f(x)$ possesses a second derivative in some neighbourhood of point c and $f^{(2)}(c) = 0$. Then, if in the indicated neighbourhood the second derivative $f^{(2)}(x)$ has unlike signs to the left and to the right of c, the graph of that function has an inflection at point $M(c, f(c))$.*

Proof. Note, first of all, that the graph of the function $y = f(x)$ has a tangent at point $M(c, f(c))$ since the hypothesis yields the existence of a finite derivative $f'(c)$. From the fact that $f^{(2)}(x)$ has unlike signs to the left and to the right of c and from Theorem 9.4 we infer that the directions of the convexity to the left and to the right of c are different. We have proved the theorem.

Example. Find points of inflection of the graph of the function $y = x^3 - 3x^2 - 4$. We have considered this function more than once (its graph is shown in Fig. 9.1). Since $f^{(2)}(x) = 6x - 6 = 6(x - 1)$, the only value of the argument for which inflection is possible is $x = 1$. This value of the argument is associated with the point of the graph $M(1, -6)$. Since $f^{(2)}(x)$ has different signs for $x > 1$ and for $x < 1$, point $M(1, -6)$ is a point of inflection of the graph of the function being considered.

9.3.3. The second sufficient condition of inflection. For the case where the investigation of the sign of the second derivative in the neighbourhood of point c is undesirable, we shall formulate the second sufficient condition of inflection presupposing the existence of a finite third derivative of the function $y = f(x)$ at point c.

Theorem 9.8. *If the function $y=f(x)$ possesses a finite third derivative at point c and satisfies the conditions $f^{(2)}(c) = 0$, $f^{(3)}(c) \neq 0$*

at that point, the graph of the function has an inflection at point $M(c, f(c))$.

Proof. It follows from the condition $f^{(3)}(c) \neq 0$ and from Theorem 8.9 that the function $f^{(2)}(x)$ either increases or decreases at point c. Since $f^{(2)}(c) = 0$, there is a neighbourhood of point c in both cases throughout which $f^{(2)}(x)$ has *different signs to the left and to the right of* c. But then, in accordance with the previous theorem, the graph of the function $y = f(x)$ has an inflection at point $M(c, f(c))$.

Remark. It is a fact that Theorem 9.8 has a narrower sphere of application than Theorem 9.7. Thus, Theorem 9.8 does not answer the question whether there is an inflection in the case where the function $y = f(x)$ possesses no finite third derivative and also in the case $f^{(3)}(c) = 0$. In the latter case, to solve the question concerning the existence of an inflection, it is necessary to investigate the behaviour at point c of derivatives of higher orders. We shall carry out this investigation in Sec. 4 of the present chapter.

We shall now return to the example discussed in the previous subsection and show that the question concerning the existence of an inflection in the graph of the function $y = x^3 - 3x^2 - 4$ can be decided with the aid of Theorem 9.8. In fact, $f^{(3)}(x) = 6 \neq 0$ and, therefore, point $M(1, -6)$ is a point of inflection in accordance with Theorem 9.8.

9.3.4. Some generalizations of the first sufficient condition of inflection. We must point out first of all that in the conditions of Theorem 9.7 we can discard the requirement that the function $y = f(x)$ must be twice differentiable *at point* c *itself*, retaining this requirement only for the points lying in some neighbourhood to the left and to the right of c. It should be additionally assumed that there exists a finite derivative $f'(c)$.

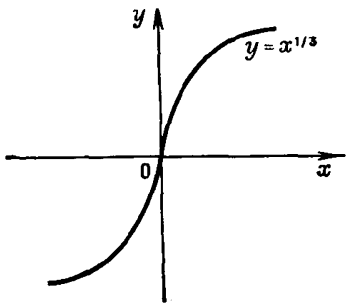


Fig. 9.11

The proof of Theorem 9.7 with the indicated alterations coincides completely with the proof given above.

When defining the point of inflection, we can agree not to exclude the case of the tangent to the graph at the point in question being *parallel* to the *Oy* axis*. With this agreement we can discard in Theorem 9.7 the requirement that the function $f(x)$ must be once differentiable at point c itself and formulate the theorem as follows.

Suppose that the function $y = f(x)$ *possesses a finite second derivative everywhere in a certain neighbourhood of point* c, *with the possible exception, may be, of point* c *itself. Also assume that the function* $y = f(x)$ *is continuous at point* c *and the graph of that function has a tangent*** *at point* $M(c, f(c))$. *Then in the case where the second derivative* $f^{(2)}(x)$ *possesses different signs within the indicated*

* This case corresponds to the infinite value of the first derivative $f'(c)$.
** At least parallel to the *Oy* axis.

neighbourhood to the left and to the right of point c, the graph of the function $y = f(x)$ *has an inflection at point* $M(c, f(c))$.

The proof of the assertion we have formulated is completely analogous to the proof of Theorem 9.7.

Example. Find the points of inflection of the graph of the function $y = x^{1/3}$. This function possesses a second derivative everywhere on an infinite straight line with the exception of point $x = 0$. The function in question is continuous at point $x = 0$, but even the first derivative turns to infinity. However, the graph of the function $y = x^{1/3}$ has a tangent at point (0,0) parallel to the Oy axis* (Fig. 9.11). Since the second derivative

$$y^{(2)} = -\frac{2}{9}\frac{1}{x^{5/3}}$$

has unlike signs to the left and to the right of point $x = 0$, the graph of the function $y = x^{1/3}$ possesses an inflection at point (0, 0).

9.4. THE THIRD SUFFICIENT CONDITION OF AN EXTREMUM AND AN INFLECTION

Theorem 9.9. Assume that $n \geqslant 1$ *is an integer and the function* $y = f(x)$ *possesses a derivative of order n in some neighbourhood of point c and a derivative of order* $n + 1$ *at point c itself. Also assume that the following relations hold true:*

$$f^{(2)}(c) = f^{(3)}(c) = \ldots = f^{(n)}(c) = 0, \quad f^{(n+1)}(c) \neq 0. \tag{9.8}$$

Then, if n is an **even** *number, the graph of the function* $y = f(x)$ *has an inflection at point* $M(c, f(c))$. *Now if n is an* **odd** *number and, besides,* $f'(c) = 0$, *then the function* $y = f(x)$ *has a local extremum at point c; to be more exact, it has a local minimum at point c for* $f^{(n+1)}(c) > 0$ *and a local maximum for* $f^{(n+1)}(c) < 0$.

Proof. (1) First suppose that n is an **even** number. At $n = 2$, the theorem being proved coincides with Theorem 9.8 we have proved, so that we need only carry out the proof for an **even** number $n \geqslant 4$.

Assume that an even number n satisfies the condition $n \geqslant 4$. It follows from the condition $f^{(n+1)}(c) \neq 0$ and from Theorem 8.9 applied to the function $f^{(n)}(x)$ that this function $f^{(n)}(x)$ either increases or decreases at point c. Since, in addition, $f^{(n)}(c) = 0$, there can be found in both cases a *sufficiently small neighbourhood of point c within which* $f^{(n)}(x)$ *has different signs to the right and to the left of c.*

Taking this fact into account we shall expand the function $f^{(2)}(x)$ in the neighbourhood of point c by Taylor's formula with the remainder in Lagrange's form. We shall see that for all x from a sufficiently small neighbourhood of point c there is a point ξ between c

* This follows, for instance, from the fact that the graph of the inverse function $x = y^3$ possesses a tangent $x = 0$ at that point.

and x such that

$$f^{(2)}(x) = f^{(2)}(c) + \frac{f^{(3)}(c)}{1!}(x-c) + \dots$$
$$\dots + \frac{f^{(n-1)}(c)}{(n-3)!}(x-c)^{n-3} + \frac{f^{(n)}(\xi)}{(n-2)!}\cdot(x-c)^{n-2}.$$

Relations 9.8 make it possible to present the last equality in the following form:

$$f^{(2)}(x) = \frac{f^{(n)}(\xi)}{(n-2)!}\cdot(x-c)^{n-2}. \tag{9.9}$$

Since within the sufficiently small neighbourhood of point c the function $f^{(n)}(x)$ has different signs at $x < c$ and $x > c$ and since ξ is always positioned between c and x, we obtain that $f^{(n)}(\xi)$, too (and the whole right-hand side of (9.9) inasmuch as n is even) possesses different signs at $x < c$ and $x > c$. But then the left-hand side of (9.9), that is $f^{(2)}(x)$, also possesses different signs in the sufficiently small neighbourhood of c at $x < c$ and $x > c$. By virtue of Theorem 9.7 this means that the graph of the function $y = f(x)$ has an inflection at point $M(c, f(c))$, and the Theorem is proved for the case of an even n.

(2) Suppose now that $n \geqslant 1$ is an odd number and it is assumed, in addition, that $f'(c) = 0$. Since at $n = 1$ the theorem we are trying to prove coincides completely with Theorem 9.2 we have proved earlier, it suffices to carry out the proof for the case of an odd $n \geqslant 3$.

Let an odd n satisfy the condition $n \geqslant 3$. For the sake of definiteness, we shall present reasons for the case $f^{(n+1)}(c) > 0$, since the proof for the case $f^{(n+1)}(c) < 0$ is analogous.

It follows from the condition $f^{(n+1)}(c) > 0$ and from Theorem 8.9 applied to the function $f^{(n)}(x)$ that this function $f^{(n)}(x)$ increases at point c. Since, besides that, $f^{(n)}(c) = 0$, this means that *there is a sufficiently small neighbourhood of point c within which $f^{(n)}(x)$ is negative to the left of c and positive to the right of c.*

Making note of this fact, we shall expand the function $f'(x)$ in the neighbourhood of point c by Taylor's formula with the remainder in Lagrange's form. We shall see that for all x from the sufficiently small neighbourhood of point c there is a point ξ between c and x such that

$$f'(x) = f'(c) + \frac{f^{(2)}(c)}{1!}\cdot(x-c) + \dots + \frac{f^{(n-1)}(c)}{(n-2)!}\cdot(x-c)^{n-2} +$$
$$+ \frac{f^{(n)}(\xi)}{(n-1)!}\cdot(x-c)^{n-1}. \tag{9.10}$$

Relations (9.8) and the additional condition $f'(c) = 0$ make it possible to rewrite equality (9.10) in the form

$$f'(x) = \frac{f^{(n)}(\xi)}{(n-1)!}\cdot(x-c)^{n-1}. \tag{9.11}$$

Since the number ξ always lies between c and x, for all x from the sufficiently small neighbourhood of c the derivative $f^{(n)}(\xi)$ is negative at $x < c$ and positive at $x > c$. When n is odd, the number $n - 1$ is even and, therefore, the whole right-hand (and, consequently, left-hand) side of (9.11) is negative to the left of c and positive to the right of c for all x from the sufficiently small neighbourhood of point c.

In accordance with Theorem 9.1 this means that the function $f(x)$ possesses a local minimum at point c. Thus for the case $f^{(n+1)}(c) > 0$

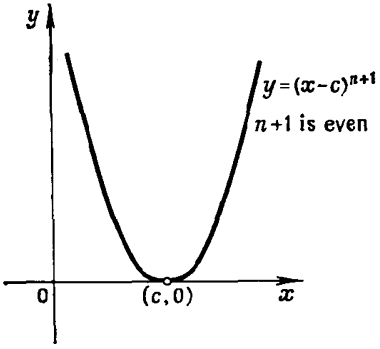


Fig. 9.12

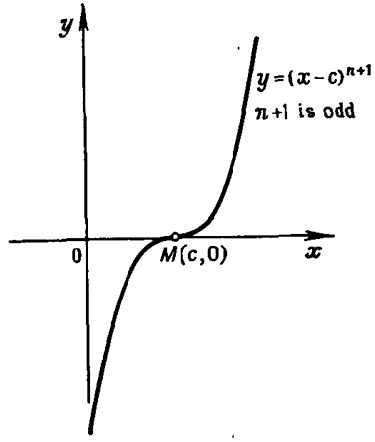


Fig. 9.13

the second part of the theorem has been proved. Since the case $f^{(n+1)}(c) < 0$ can be treated by a complete analogy, we have proved the theorem.

Example. Find out whether the function $f(x) = (x - c)^{(n+1)}$ has an extremum and an inflection. It is easy to see that $f'(c) = f^{(2)}(c) = \ldots = f^{(n)}(c) = 0$, $f^{(n+1)}(c) = (n+1)! > 0$. In accordance with Theorem 9.9 the function possesses a minimum at point $x = c$ for an *even* $(n + 1)$ (Fig. 9.12), while for an *odd* $(n + 1)$ the graph of the function possesses an inflection at point $M(c, 0)$ (Fig. 9.13).

9.5. ASYMPTOTES OF THE GRAPH OF A FUNCTION

Definition 1. *A straight line $x = a$ is said to be a* **vertical asymptote** *of the graph of the function $y = f(x)$ if at least one of the limiting values*

$$\lim_{x \to a+0} f(x) \text{ or } \lim_{x \to a-0} f(x)$$

is equal to $+\infty$ or $-\infty$.

Example. The graph of the function $y=\frac{1}{x}$ has a vertical asymptote $x=0$, for $\lim\limits_{x\to 0+0}\frac{1}{x}=+\infty$, $\lim\limits_{x\to 0-0}\frac{1}{x}=-\infty$ (Fig. 9.14).

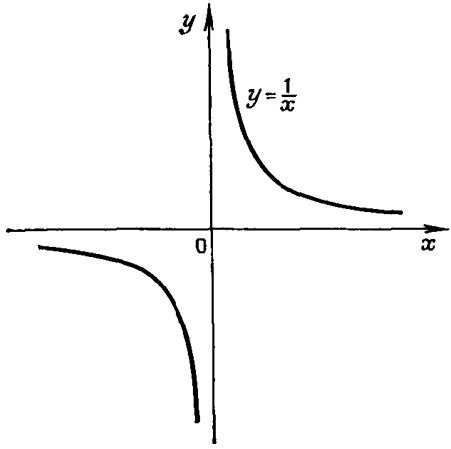


Fig. 9.14

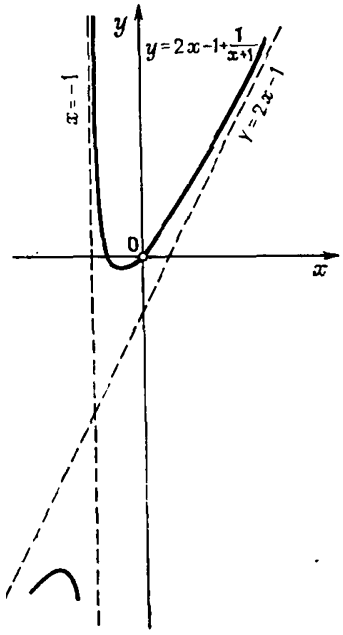


Fig. 9.15

Let us further postulate that the function $y=f(x)$ is defined for arbitrarily large values of the argument. For definiteness, we shall consider arbitrarily large values of a *positive* sign.

Definition 2. *The straight line*

$$Y=kx+b \tag{9.12}$$

is said to be an **inclined asymptote** *of the graph of the function* $y=f(x)$ *for* $x\to+\infty$ *if the function* $f(x)$ *can be represented in the form*

$$f(x)=kx+b+\alpha(x), \tag{9.13}$$

where $\lim\limits_{x\to+\infty}\alpha(x)=0$.

Theorem 9.10. *For the graph of the function* $y=f(x)$ *to have an inclined asymptote (9.12) for* $x\to+\infty$, *it is necessary and sufficient that there exist two limiting values*

$$\lim_{x\to+\infty}\frac{f(x)}{x}=k \text{ and } \lim_{x\to+\infty}[f(x)-kx]=b. \tag{9.14}$$

Proof. (1) *Necessity.* Assume that the graph of the function $y = f(x)$ possesses, for $x \to +\infty$, an asymptote (9.12), that is, representation (9.13) holds for $f(x)$. Then

$$\lim_{x\to+\infty} \frac{f(x)}{x} = \lim_{x\to+\infty} \frac{kx+b+\alpha(x)}{x} = \lim_{x\to+\infty} \left[k + \frac{b}{x} + \frac{\alpha(x)}{x}\right] = k,$$

$$\lim_{x\to+\infty} [f(x) - kx] = \lim_{x\to+\infty} [b + \alpha(x)] = b.$$

(2) *Sufficiency.* Suppose there exists a limiting value (9.14). The second of these limiting values enables us to assert that the difference $f(x) - kx - b$ is infinitesimal for $x \to +\infty$. Designating this infinitesimal quantity as $\alpha(x)$, we obtain for $f(x)$ representation (9.13) and thus prove the theorem.

Remark. The definition of an inclined asymptote and the proof of Theorem 9.10 are the same for the case $x \to -\infty$.

Example. The graph of the function $y = \frac{2x^2+x}{x+1} = 2x - 1 + \frac{1}{x+1}$ has an inclined asymptote $Y = 2x - 1$ both for $x \to +\infty$ and for $x \to -\infty$ and, besides, has a vertical asymptote $x = -1$ (Fig. 9.15). Indeed,

$$\lim_{x\to\pm\infty} \frac{f(x)}{x} = \lim_{x\to\pm\infty} \frac{2x^2+x}{x(x+1)} = 2,$$

$$\lim_{x\to\pm\infty} [f(x) - 2x] = \lim_{x\to\pm\infty} \left[-1 + \frac{1}{x+1}\right] = -1,$$

$$\lim_{x\to-1+0} f(x) = +\infty, \quad \lim_{x\to-1-0} f(x) = -\infty.$$

Besides a linear asymptote (9.12) there exist asymptotes of a more intricate shape.

It is said that a parabola of order n specified by the polynomial

$$Y = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 \qquad (9.12^*)$$

is an asymptote of the graph of the function $y = f(x)$ for $x \to +\infty$ if the function $f(x)$ can be represented in the form

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 + \alpha(x),$$

where $\lim_{x\to+\infty} \alpha(x) = 0$.

The following assertion can easily be proved.

For the graph of the function $y = f(x)$ to have an asymptote (9.12) for $x \to +\infty$, it is necessary and sufficient that there exist the following $n + 1$ limiting values:*

$$\lim_{x\to+\infty} \frac{f(x)}{x^n} = a_n, \quad \lim_{x\to+\infty} \frac{f(x) - a_n x^n}{x^{n-1}} = a_{n-1}, \ldots,$$

$$\lim_{x\to+\infty} \frac{f(x) - (a_n x^n + a_{n-1}x^{n-1} + \ldots + a_2 x^2)}{x} = a_1,$$

$$\lim_{x\to+\infty} [f(x) - (a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x)] = a_0.$$

9.6. THE SCHEME FOR INVESTIGATING THE GRAPH OF A FUNCTION

In this section we shall elucidate the scheme most expedient for investigating the graph of a function and give an example illustrating the scheme.

The following investigation should be carried out to make a qualitative estimation of the graph of the function $y = f(x)$.

1°. Determine the domain of definition of the function.

2°. Elucidate the question of the existence of asymptotes (vertical and inclined).

3°. Find the domains of an increase and a decrease of the function and the extremal points.

4°. Determine the domain where the direction of the convexity is retained and find the inflection points.

5°. Find the points of intersection of the graph of the function and the Ox axis.

Proceeding from the data obtained, it is easy to make a rough drawing of the graph of the function. By way of illustration, let us construct the graph of the function

$$y = \frac{2x^3 - 5x^2 + 14x - 6}{4x^2}. \tag{9.15}$$

We shall follow the scheme given above.

1°. Since the function (9.15) is a rational fraction, it is defined and continuous everywhere on an infinite line, except for the point $x = 0$ where the denominator turns into zero.

2°. As to the existence of asymptotes, it is evident that

$$\lim_{x \to 0 \pm 0} \frac{2x^3 - 5x^2 + 14x - 6}{4x^2} = -\infty,$$

and, therefore, the graph of the function possesses a *vertical asymptote* $x = 0$. Next, from the existence of the limiting values

$$\lim_{x \to \pm\infty} \frac{f(x)}{x} = \lim_{x \to \pm\infty} \frac{2x^3 - 5x^2 + 14x - 6}{4x^3} =$$

$$= \lim_{x \to \pm\infty} \frac{2 - \frac{5}{x} + \frac{14}{x^2} - \frac{6}{x^3}}{4} = \frac{1}{2},$$

$$\lim_{x \to \pm\infty} \left[f(x) - \frac{x}{2} \right] = \lim_{x \to \pm\infty} \frac{2x^3 - 5x^2 + 14x - 6 - 2x^3}{4x^2} =$$

$$= \lim_{x \to \pm\infty} \frac{-5 + \frac{14}{x} - \frac{6}{x^2}}{4} = -\frac{5}{4}$$

it follows that both for $x \to +\infty$ and $x \to -\infty$ the graph of the function possesses an *inclined asymptote* $Y = \frac{x}{2} - \frac{5}{4}$.

3°. To determine the domains of an increase and a decrease of the function, let us calculate the first derivative of the function (9.15):

$$y' = \frac{x^3 - 7x + 6}{2x^3} = \frac{(x-1)(x-2)(x+3)}{2x^3}.$$

Also bearing in mind that for $x = 0$ neither the function itself nor its first derivative exists, we obtain the following domains where the sign of y' is retained:

Domain of the values of x	$-\infty < x < -3$	$-3 < x < 0$	$0 < x < 1$	$1 < x < 2$	$2 < x < \infty$
The sign of y'	+	−	+	−	+
Behaviour of the function	increases	decreases	increases	decreases	increases

It can be seen from the table that the function has the following extremal points:

(1) maximum at $x = -3$, with $f(-3) = -\frac{49}{12}$,

(2) maximum at $x = 1$, with $f(1) = \frac{5}{4}$,

(3) minimum at $x = 2$, with $f(2) = \frac{9}{8}$.

4°. To determine the domains where the direction of convexity is retained we must calculate the second derivative:

$$y^{(2)} = \frac{7x - 9}{x^4} = \frac{7\left(x - \frac{9}{7}\right)}{x^4}.$$

Bearing in mind that neither the function itself nor its derivatives exist at point $x = 0$, we obtain the following domains where the sign of $y^{(2)}$ is retained:

Domain of the values of x	$-\infty<x<0$	$0<x<\frac{9}{7}$	$\frac{9}{7}<x<\infty$
Sign of $y^{(2)}$	$-$	$-$	$+$
Direction of convexity of the graph	upwards	upwards	downwards

The data in the table show that the graph of the function has an inflection point $\left(\frac{9}{7}, f\left(\frac{9}{7}\right)\right)$. It is easy to compute that $f\left(\frac{9}{7}\right)=\frac{913}{756}$.

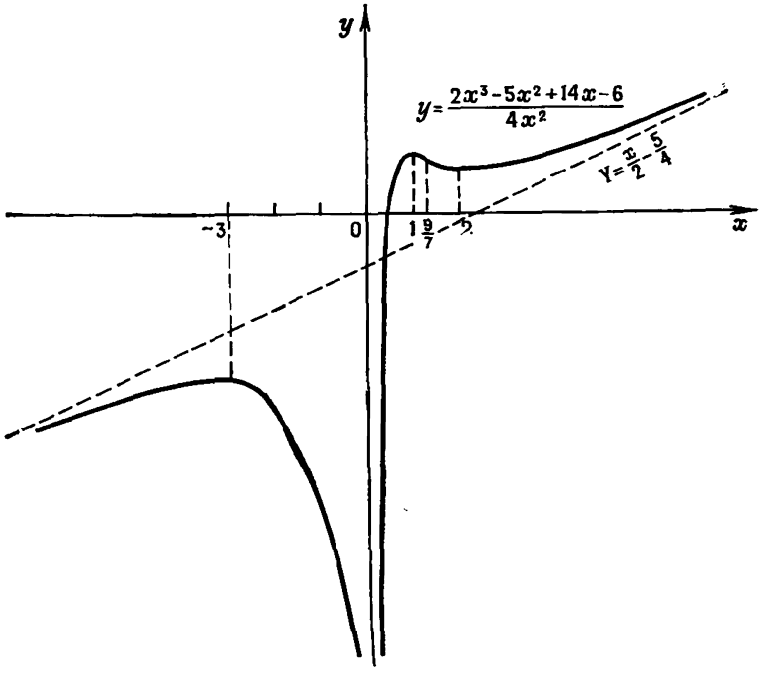


Fig. 9.16

5°. It remains to determine the points of intersection of the graph with the Ox axis. These points correspond to the real roots of the equation

$$2x^3 - 5x^2 + 14x - 6 = 0.$$

It is easy to see that $2x^3-5x^2+14x-6=2\left(x-\frac{1}{2}\right)(x^2-2x+6)$. Since the quadratic trinomial (x^2-2x+6) has complex roots,

the equation being considered has only one real root $x=\frac{1}{2}$, so that the graph of the function intersects the x-axis at point $\left(\frac{1}{2}, 0\right)$. These data allow us to make a rough drawing of the graph of the function under consideration (Fig. 9.16).

9.7. FINDING THE MAXIMUM AND THE MINIMUM VALUE OF A FUNCTION. AN END-POINT EXTREMUM

9.7.1. Finding the maximum and minimum value of a function. Let us consider a function $y = f(x)$ defined and continuous on the closed interval $[a, b]$. Up till now we were interested only in local

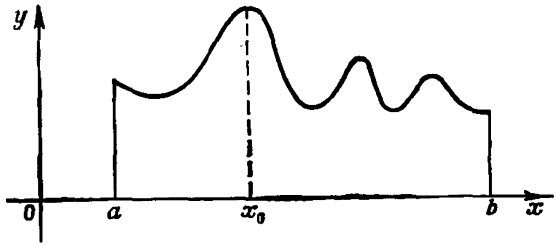


Fig. 9.17

maxima and minima of this function, and now we shall try *to find the maximum and minimum values* of $f(x)$ on the interval $[a, b]$. We want to emphasize the fact that in accordance with Weierstrass's theorem (see 8.6), the function $f(x)$ is sure to reach its maximum (minimum) value at a certain point of the interval $[a, b]$. For definiteness, let us restrict ourselves to finding the maximum value of $f(x)$ on the interval $[a, b]$.

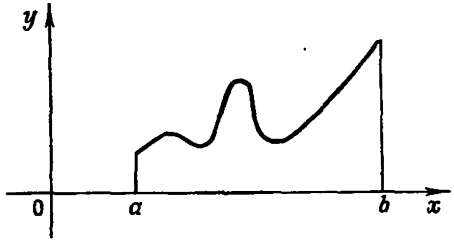


Fig. 9.18

The function $f(x)$ can reach its maximum value either at an interior point x_0 of the interval $[a, b]$ (then it coincides with one of the local maxima of the function $f(x)$, see Fig. 9.17), or at one of the end points of the interval $[a, b]$ (Fig. 9.18). Hence it is clear that to find the maximum value of the function $f(x)$ on the interval $[a, b]$ it is necessary to compare the values of $f(x)$ at all the points of the local maximum and at the end points a and b of the closed interval. The greatest of these values will be the maximum value of $f(x)$ on the interval $[a, b]$. The minimum value of $f(x)$ on the interval $[a, b]$ can be found by analogy.

If we want to obviate the investigation of the points of a possible extremum, we can simply compare the values of $f(x)$ at all the points of a possible extremum and at the end points a and b. The greatest (least) value will, evidently, be the maximum (minimum) value of the function $f(x)$ on the interval $[a, b]$.

Make note of the fact that if on the closed interval $[a, b]$ $f(x)$ possesses *only one* point of a local extremum* which is the point of a local maximum (minimum), then, without comparing the value

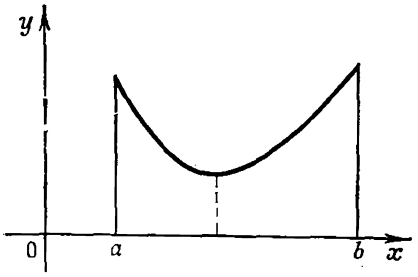


Fig. 9.19

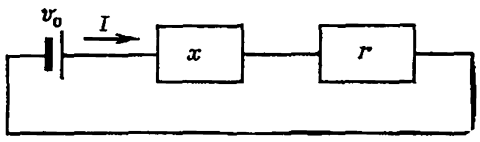


Fig. 9.20

of $f(x)$ at that point with the values $f(a)$ and $f(b)$, we can state that this value is the maximum (minimum) value of $f(x)$ on the interval $[a, b]$ (Fig. 9.19).

In a similar way we can solve the problem of finding the maximum (and minimum) value of the function $y = f(x)$ on an interval, half-line and an infinite line (under the condition that this value does exist).

It may so happen that the function $f(x)$ does not have points of a possible extremum on the closed interval $[a, b]$ (or a half-line $a \leqslant x < \infty$). In that case $f(x)$ is monotonic throughout the closed interval (the half-line) and its maximum and minimum values are reached at the end points of that interval (at the end points of the half-line). We shall illustrate the latter case by an example from physics. Suppose it is required to determine what resistance x should be connected to the circuit in series with the given resistance r to develop on r the greatest power possible (the voltage v_0 of the battery is taken to be constant, see Fig. 9.20). In accordance with Ohm's law, the current I in the circuit is equal to

$$I = \frac{v_0}{r+x}.$$

Consequently, according to the same law, the voltage drop v_r at the resistance r is equal to

$$v_r = Ir = \frac{v_0 r}{r+x}.$$

* Precisely such a case is often encountered in practice.

Thus, the power $w(x)$ developed on the resistance r is

$$w(x) = Iv_r = \frac{v_0^2 r}{(r+x)^2}.$$

Since from the point of view of physics the resistance x cannot be negative, the problem reduces to searching for the greatest value of the function $w(x)$ on the half-line $x \geqslant 0$. Having calculated the derivative of this function

$$w'(x) = -\frac{2v_0^2 r}{(r+x)^3},$$

we shall verify that $w'(x) < 0$ everywhere on the half-line $x \geqslant 0$ and there are no points of a possible extremum. Thus the function

Fig. 9.21

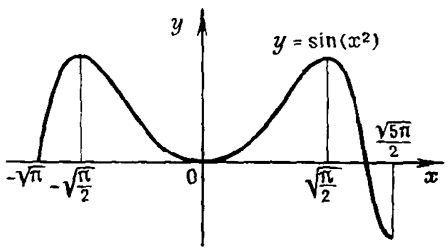


Fig. 9.22

$w(x)$ decreases everywhere on the half-line $x \geqslant 0$ and its maximum value on that half-line can be reached at $x = 0$ and is equal to v_0^2/r (Fig. 9.21). This is also clear from physical considerations.

Let us consider another example, that of searching for the maximum and minimum values of the function $y = \sin(x^2)$ on the closed interval $-\sqrt{\pi} \leqslant x \leqslant \frac{\sqrt{5\pi}}{2}$.

Since $y' = 2x\cos x^2$, the indicated function possesses three points of a possible extremum on the interval being considered, $x = 0$ and $x = \pm\sqrt{\frac{\pi}{2}}$. Comparing the values of the function at the indicated points and at the end points of the interval

$$f(0) = 0, \quad f\left(\pm\sqrt{\frac{\pi}{2}}\right) = 1, \quad f\left(-\sqrt{\pi}\right) = 0,$$

$$f\left(\frac{\sqrt{5\pi}}{2}\right) = \sin\frac{5\pi}{4} = -\frac{\sqrt{2}}{2},$$

we shall make sure that the maximum value of the function under consideration is equal to $+1$ and is reached at two interior points of the interval, $x_1 = -\sqrt{\frac{\pi}{2}}$ and $x_2 = +\sqrt{\frac{\pi}{2}}$,

and the minimum value of the function being considered is equal to $-\frac{\sqrt{2}}{2}$ and is reached at the right end point of the interval, $\frac{\sqrt{5\pi}}{2}$.

The graph of this function is illustrated in Fig. 9.22.

9.7.2. End-point extremum. Assume that the function $y=f(x)$ is defined on a certain closed interval $[a, b]$. We shall say that the function possesses an *end-point maximum (end-point minimum)* at point b of the interval if there is a left-hand semi-neighbourhood of point b throughout which the value $f(b)$ is the greatest (the least) among all the values of the function. We can use the same method to determine the end-point maximum and end-point minimum at the boundary point a of the interval $[a, b]$. The end-point maximum and the end-point minimum go under the common term the *end-point extremum*. We can state the following *sufficient condition for the end-point extremum: for the function $y = f(x)$ to have an end-point maximum (end-point minimum) at point b of the interval $[a, b]$, it is sufficient that this function should possess at a point b a positive (negative) left-hand derivative**. (The proof is analogous to that of Theorem 8.9.) The indicated sufficient condition for the end-point extremum immediately yields the following *necessary condition for the end-point extremum of the function possessing a left-hand derivative at point b: for the function $y = f(x)$ possessing a left-hand derivative at point b to have an end-point maximum (end-point minimum) at that point, it is necessary that the indicated derivative be nonnegative (nonpositive).*

In conclusion we shall prove the following remarkable assertion.

Theorem 9.11 (Darboux's theorem**). *Assume that the function $f(x)$ possesses a finite derivative everywhere on the closed interval $[a,b]$***, and suppose that $f'(a+0) = A$, $f'(b-0) = B$. Then, whatever the number C contained between A and B, there is a point ξ on that interval such that $f'(\xi) = C$****.*

Proof. We shall first prove the following assertion: if $F(x)$ possesses a finite derivative on $[a, b]$ and if $F'(a+0)$ and $F'(b-0)$ are numbers of unlike signs, there can be found a point ξ on the interval $[a, b]$ such that $F'(\xi) = 0$.

Suppose for the sake of definiteness that $F'(a+0) < 0$, $F'(b-0) > 0$. Then, the function $F(x)$ possesses an end-point maximum at both end points of the interval $[a, b]$. But this means that the minimum value of $F(x)$ on the interval $[a, b]$ can be reached at some interior point ξ of that interval (the function $F(x)$ is differentiable and, hence, continuous on the interval $[a, b]$ and,

---

* For the boundary point a the sufficient condition for the end-point maximum (end-point minimum) is the negativity (positivity) of the right-hand derivative at point a.

** Gaston Darboux (1842-1917), a French mathematician.

*** The function $f(x)$ is understood here to have a derivative at any interior point of the closed interval $[a, b]$ and, besides, to have a left-hand derivative at point b and a right-hand derivative at point a.

**** Make note of the fact that the continuity of the derivative $f'(x)$ is not assumed here.

consequently, reaches its minimum value on that interval). At the indicated point ξ the function $F(x)$ has a local minimum and, therefore, $F'(\xi) = 0$.

To prove Theorem 9.11, it remains to put $F(x) = f(x) - Cx$ and apply the assertion we have just proved to $F(x)$.

Remark. If the derivative $f'(x)$ exists everywhere in some neighbourhood of point c, it cannot possess a discontinuity of the first kind or a removable discontinuity at point c. This assertion, which we have proved in 8.10.3, follows immediately from Darboux's theorem.

CHAPTER 10

DEFINITE INTEGRAL

In Chapter 1 we considered the physical problem of calculating the path traversed by a particle moving along the Oy axis proceeding from the known velocity of the particle and the geometrical problem of calculating the area of a *curvilinear trapezoid* (that is the figure lying between the graph of the function $y = f(x)$ and the closed interval $[a, b]$ of the Ox-axis). The consideration of these two problems naturally led us to the necessity of introducing in Chapter 1 a new mathematical notion, that of the *definite integral.* Besides the two problems we have studied, a number of other important physical and geometrical problems lead us to the concept of the definite integral. The present chapter is devoted to the theory of the definite integral and the next chapter contains information on the application of the theory to some problems from geometry and physics.

10.1. INTEGRAL SUMS. INTEGRABILITY

Assume that a function $f(x)$ is given on an interval $[a, b]$, $a < b$. Let us denote by T the partition of the interval $[a, b]$ by some non-coinciding points $a = x_0 < x_1 < \ldots < x_n = b$ into n subintervals $[x_0, x_1], [x_1, x_2], \ldots, [x_{n-1}, x_n]$. We shall call the points $x_0, x_1, \ldots, x_n$ the points of partition T. Suppose ξ_i is an arbitrary point of the subinterval $[x_{i-1}, x_i]$ and Δx_i is the difference $x_i - x_{i-1}$ which we shall be calling the length of the subinterval $[x_{i-1}, x_i]$.

Definition 1. *The number* $I\{x_i, \xi_i\}$ *where*

$$I\{x, \xi_i\} = f(\xi_1)\,\Delta x_1 + f(\xi_2)\,\Delta x_2 + \ldots$$

$$\ldots + f(\xi_n)\,\Delta x_n = \sum_{i=1}^{n} f(\xi_i)\,\Delta x_i$$

is known as the **integral sum** *of the function* $f(x)$ *corresponding to the given partition* T *of the interval* $[a, b]$ *and to the given choice of intermediate points* ξ_i *on the subintervals,* $[x_{i-1}, x_i]$. In what follows we shall use Δ to denote the **length of the maximum subinterval** of partition T, i.e. $\Delta = \max \Delta x_i$.

Let us elucidate the geometrical meaning of an integral sum. For that purpose we shall consider a *curvilinear trapezoid*, that is, a figure bounded by the graph of the function $f(x)$ (for simplicity we shall consider the function to be positive and continuous), by two ordinates drawn through points a and b of the abscissa axis and by the axis of abscissae (Fig.10.1). The integral sum $I\{x_i, \xi_i\}$ evidently represents the area of the stepped figure hatched in Fig. 10.1.

Definition 2. *The number I is called the* **limit of the integral sums** $I\{x_i, \xi_i\}$ *for* $\Delta \to 0$ *if for any positive number* ε *we can indicate a*

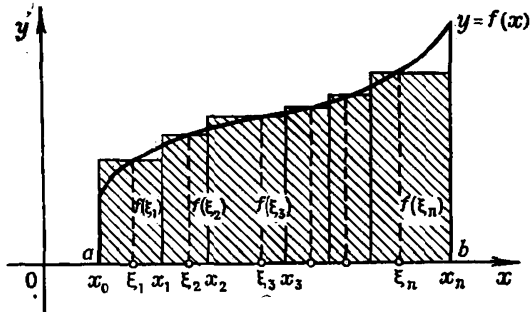


Fig. 10.1

positive number δ* *such that for any partition T of the interval* $[a, b]$, *the maximum length* Δ *of whose subintervals is less than* δ, *the inequality*

$$|I\{x_i, \xi_i\} - I| < \varepsilon$$

holds true irrespective of the choice of points ξ_i *on the intervals* $[x_{i-1}, x_i]$.

To designate the limit of integral sums, use is usually made of the notation

$$I = \lim_{\Delta \to 0} I\{x_i, \xi_i\}.$$

Definition 3. *The function* $f(x)$ *is said to be* **integrable** (*Riemann*** *integrable*) *on the closed interval* $[a, b]$ *if there exists a finite limit I of the integral sums of that function for* $\Delta \to 0$. *The indicated limit I is called the definite integral of the function* $f(x)$ *with respect to the interval* $[a, b]$ *and is designated as follows*:

$$I = \int_a^b f(x)\, dx.$$

---

* Since the number δ depends on ε they sometimes write $\delta = \delta(\varepsilon)$.

** Bernhard Riemann (1826-1866), a German mathematician.

Visual geometric representations show* that the definite integral is numerically equal to the area of a curvilinear trapezoid defined by the graph of the function $f(x)$ on the interval $[a, b]$. The validity of this assertion will be proved in Chapter 11.

Here is an example of an *integrable function*. We shall prove that the function $f(x)=c=\text{const}$ is integrable on any interval $[a, b]$ with $\int_a^b c\,dx=c(b-a)$. Indeed, since $f(\xi_i)=c$ for any ξ_i, we have

$$I\{x_i, \xi_i\} = c\Delta x_1 + c\Delta x_2 + \ldots + c\Delta x_n = c(\Delta x_1 + \Delta x_2 + \ldots + \Delta x_n) = c(b-a);$$

consequently we have $\lim\limits_{\Delta\to 0} I\{x_i, \xi_i\} = c(b-a)$.

Let us consider the question of integrability of functions unlimited on the interval $[a, b]$.

We shall prove the following assertion: *the function $f(x)$ unlimited on the interval $[a, b]$ is nonintegrable on that interval.*

Proof. Assume that the function $f(x)$ is unlimited on the interval $[a, b]$. Then is it unlimited on a certain subinterval $[x_{k-1}, x_k]$ of *any* given partition T of the interval $[a, b]$. Therefore, the summand $f(\xi_k)\Delta x_k$ of the integral sum $I\{x_i, \xi_i\}$ corresponding to that partition T can be made arbitrarily large in its absolute value due to the choice of point ξ_k. From this it follows that the integral sums $I\{x_i, \xi_i\}$ corresponding to any partition T are unlimited** and, therefore, there is no finite limit of integral sums.

Taking into consideration the assertion we have proved, we shall only discuss the functions limited on the interval $[a, b]$. The question arises of *whether every function limited on the interval $[a, b]$ is integrable on that interval.* The following example shows that in the general case it is not so. Let us verify that the *Dirichlet function*, which is obviously limited on the interval $[a, b]$ and whose values are equal to unity at rational points and to zero at irrational points, is *not integrable* on the interval $[a, b]$. Indeed, if for *any* partition T *with arbitrarily small* Δ we assume points ξ_i to be rational, then, evidently, $I\{x_i, \xi_i\}=\sum\limits_{i=1}^{n} f(\xi_i)\Delta x_i=\sum\limits_{i=1}^{n}\Delta x_i=b-a$. Now if for the same partition T we assume points ξ_i to be irrational, then

---

* See 1.4.

** To verify this, it is sufficient to fix points ξ_i on all the subintervals of partition T, except for the interval $[x_{k-1}, x_k]$. Then only the summand $f(\xi_k)\Delta x_k$ will vary in the integral sum $I\{x_i, \xi_i\}$, which can be arbitrarily large in its absolute value.

$I\{x_i, \xi_i\} = 0$. Therefore, there is no limit of integral sums for the Dirichlet function, that is, the function is not integrable.

Later on we shall prove integrability of all continuous functions and of a wide range of discontinuous functions.

10.2. UPPER AND LOWER SUMS

10.2.1. The concept of an upper and a lower sum. Let the function $f(x)$ be bounded on the closed interval $[a, b]$ and T be a partition of the interval by points $a = x_0 < x_1 < \ldots < x_n = b$. Let us denote

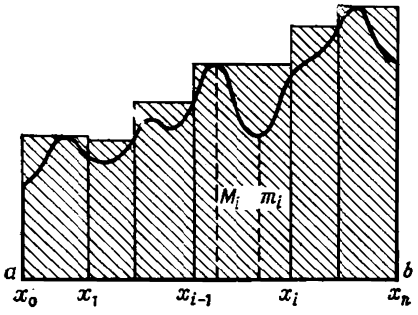


Fig. 10.2

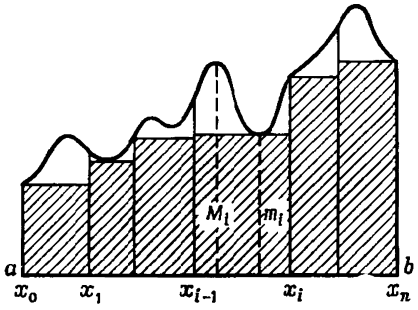


Fig. 10.3

by M_i and m_i, respectively, the least upper bound and the greatest lower bound of that function on the interval $[x_{i-1}, x_i]$. *The sums*

$$S = M_1 \Delta x_1 + M_2 \Delta x_2 + \ldots + M_n \Delta x_n = \sum_{i=1}^{n} M_i \Delta x_i$$

and

$$s = m_1 \Delta x_1 + m_2 \Delta x_2 + \ldots + m_n \Delta x_n = \sum_{i=1}^{n} m_i \Delta x_i$$

are called the **upper** *and the* **lower sum**, *respectively, of the function* $f(x)$ *for the given partition* T *of the interval* $[a, b]$.

It is evident that *any integral sum* $I\{x_i, \xi_i\}$ *of the given partition* T *of the interval* $[a, b]$ *is contained between the upper sum* S *and the lower sum* s *of that partition.*

The concepts of the *upper* and *lower* sums become especially clear if we make use of geometric representation. For the sake of simplicity, we shall consider a positive and continuous function $f(x)$ and a curvilinear trapezoid defined by that function (Figs. 10.2 and 10.3). If T is a partition of the interval $[a, b]$, then the numbers M_i and m_i are, in the case of the continuous function $f(x)$, the maximum and the minimum values of that function on the subinterval $[x_{i-1}, x_i]$ of partition T. Therefore, the upper sum S is equal to the area of the

stepped figure hatched in Fig. 10.2. which *contains* the curvilinear trapezoid while the lower sum s is equal to the area of the stepped figure hatched in Fig. 10.3 which *is contained* in the curvilinear trapezoid (the trapezoid is shown in bold line in Figs. 10.2 and 10.3).

As has been pointed out, it is clear from geometric interpretation that the integral is numerically equal to the area of the curvilinear trapezoid. It is evident, on the other hand, that if the difference between the upper and the lower sum can be made arbitrarily small, then these sums may become arbitrarily close to the area of the curvilinear trapezoid, and, therefore, it may be hoped that for the function to be integrable it is necessary and sufficient that the difference between the upper and the lower sum could be arbitrarily small. A rigorous proof of this assertion will be given in the next section.

10.2.2. The properties of the upper and lower sums. Let us prove the validity of the following properties of the upper and lower sums:

1°. *For any fixed partition T and for any positive ε we can choose intermediate points ξ_i on the intervals $[x_{i-1}, x_i]$ so that the integral sum $I\{x_i, \xi_i\}$ will satisfy the inequalities $0 \leqslant S - I\{x_i, \xi_i\} < \varepsilon$. The points ξ_i may also be chosen so that the integral sum will satisfy the inequalities $0 \leqslant I\{x_i, \xi_i\} - s < \varepsilon$.*

Suppose that T is some fixed partition of the closed interval $[a, b]$. We shall prove, for example, the possibility of choosing, for a given positive ε, points ξ_i so that the inequality $0 \leqslant S - I\{x_i, \xi_i\} < \varepsilon$ will hold true. According to the definition of the least upper bound M_i, there can be chosen, for the given $\varepsilon > 0$, a point ξ_i on the interval $[x_{i-1}, x_i]$ such that

$$0 \leqslant M_i - f(\xi_i) < \frac{\varepsilon}{b-a}, \quad i = 1, 2, \ldots, n.$$

Multiplying these inequalities by Δx_i and adding up the results, we shall have

$$0 \leqslant S - I\{x_i, \xi_i\} < \varepsilon.$$

We have thus established the validity of property 1°.

2°. *If partition T' of the interval $[a, b]$ has been obtained by adding new points to the points of partition T of that interval, then the upper sum S' of partition T' is not greater than the upper sum S of partition T and the lower sum s' of partition T' is not less than the lower sum s of partition T, that is*

$$s \leqslant s', \quad S' \leqslant S.$$

Since partition T' can be obtained from partition T by successive addition of new points to the latter, it is, evidently, sufficient to prove the formulated property for the case where one point is added to partition T. Let this new point x' be located on the interval $[x_{i-1}, x_i]$ of partition T of the interval $[a, b]$. We shall designate by

M_i' and M_i'' the least upper bounds of the function $f(x)$ on the intervals $[x_{i-1}, x']$ and $[x', x_i]$, by $\Delta x_i'$ and $\Delta x_i''$ the lengths of these intervals, and by S and S' the upper sums of partition T and partition T' obtained by adding point x' to partition T. Make note of the fact that $\Delta x_i = \Delta x_i' + \Delta x_i''$. Moreover, if M_i is the least upper bound of the values of the function $f(x)$ on the closed interval $[x_{i-1}, x_i]$, then $M_i \geqslant M_i'$ and $M_i \geqslant M_i''$ since it is evident that *the least upper bound of the function on a portion of the interval* $[x_{i-1}, x_i]$ *does not exceed the least upper bound* M_i *of that function on the entire interval* $[x_{i-1}, x_i]$. Taking into consideration, therefore, that the sums S and S' differ only by the summands $M_i \Delta x_i$ and $M_i' \Delta x_i' + M_i'' \Delta x_i''$, we obtain

$$S - S' = M_i \Delta x_i - (M_i' \Delta x_i' + M_i'' \Delta x_i'') =$$
$$= (M_i - M_i') \Delta x_i' + (M_i - M_i'') \Delta x_i'' \geqslant 0,$$

that is $S' \leqslant S$. The proof for the lower sums is quite analogous.

3°. *Assume that* T' *and* T'' *are any two partitions of the closed interval* $[a, b]$. *Then the lower sum of one of these partitions does not exceed the upper sum of the other, namely, if* s', S' *and* s'', S'' *are the lower and the upper sums of partitions* T' *and* T'' *respectively, then*

$$s' \leqslant S'' \text{ and } s'' \leqslant S'.$$

We have earlier established that the lower sum of the given partition does not exceed the upper sum of that partition. Suppose that T is a partition of the closed interval $[a, b]$ obtained by combining partitions* T' and T'', and s and S are the upper and the lower sum of partition T. Since partition T can be obtained from partition T' by adding points of partition T'' to it, property 2° and the indicated property of the upper and lower sums of the same partition yield

$$s' \leqslant s \leqslant S \leqslant S'.$$

However, partition T can also be obtained from partition T'' by adding points of partition T' to it. Therefore,

$$s'' \leqslant s \leqslant S \leqslant S''.$$

Comparing the inequalities obtained above with the latter inequalities, we see that $s' \leqslant S''$, $s'' \leqslant S'$.

We have established the validity of property 3°.

4°. *The set* $\{S\}$ *of the upper sums of the given function* $f(x)$ *for various partitions of the closed interval* $[a, b]$ *is bounded below. The set* $\{s\}$ *of the lower sums is bounded above.*

This property follows immediately from property 3°. Indeed, any upper sum is not smaller than some fixed lower sum. Consequently,

* In this case the common points of partitions T' and T'' are taken into account only once.

the set $\{S\}$ of the upper sums is bounded below. Any lower sum does not exceed any one of the upper sums and, therefore, the set $\{s\}$ of the lower sums is bounded above. Let us denote by $\overline{I}$ the greatest lower bound (supremum) of the set $\{S\}$ of upper sums and by $\underline{I}$ the least upper bound (infimum) of the set of lower sums:

$$\overline{I} = \inf\{S\}, \quad \underline{I} = \sup\{s\}.$$

The numbers $\overline{I}$ and $\underline{I}$ are called *the upper and the lower Darboux integrals of the function* $f(x)$ respectively. Let us prove that $\underline{I} \leqslant \overline{I}$. Assume $\underline{I} > \overline{I}$. Then the difference $\underline{I} - \overline{I}$ is a positive number which we shall designate by ε so that $\underline{I} - \overline{I} = \varepsilon > 0$. It follows from the definition of the infimum $\underline{I}$ and the supremum $\overline{I}$ that there are numbers S' and s'' representing the upper and the lower sums, respectively, of partitions T' and T'' of the interval $[a, b]$ such that $\overline{I} + \frac{\varepsilon}{2} > S'$ and $\underline{I} - \frac{\varepsilon}{2} < s''$. Subtracting the second inequality from the first and taking into account that $\underline{I} - \overline{I} = \varepsilon$, we obtain $s'' > S'$. But this last inequality contradicts property 3° of the upper and lower sums.

5°. *Suppose that partition T' of the interval $[a, b]$ has been obtained from partition T by adding p new points to the latter and assume that s', S' and s, S are the lower and the upper sums, respectively, of partitions T' and T. Then, for the differences $S - S'$ and $s' - s$* we can obtain an evaluation depending on the maximum length Δ of the subintervals of partition T, on the number p of the added points and on the least upper bound M and the greatest lower bound m of the function $f(x)$ on the interval $[a, b]$, namely,*

$$S - S' \leqslant (M - m)\, p\Delta, \quad s' - s \leqslant (M - m)\, p\Delta.$$

To ascertain the validity of this property, it suffices to prove the indicated two inequalities for the case where only one point, x', is added to partition T. Suppose this point is on the interval $[x_{i-1}, x_i]$ of partition T. Then it will divide the interval into two intervals, $[x_{i-1}, x']$ and $[x', x_i]$ whose lengths will be denoted by $\Delta x'_i$ and $\Delta x''_i$, respectively. Assume that M'_i, M'_i and M''_i are the least upper bounds of the function $f(x)$ on the intervals $[x_{i-1}, x_i]$, $[x_{i-1}, x']$ and $[x', x_i]$, respectively. Since $\Delta x_i = \Delta x'_i + \Delta x''_i$ and the upper sums S and S' of partitions T and T' differ only by the summands $M_i \Delta x_i$ and $M'_i \Delta x'_i + M''_i \Delta x''_i$, we have $S - S' = M_i (\Delta x'_i + \Delta x''_i) - (M'_i \Delta x'_i + M''_i \Delta x''_i) = (M_i - M'_i)\, \Delta x'_i + (M_i - M''_i)\, \Delta x''_i$. Furthermore,

* We want to point out that by virtue of property 2° these differences are nonnegative.

$m \leqslant M_i' \leqslant M_i \leqslant M$ and $m \leqslant M_i'' \leqslant M_i \leqslant M^*$, therefore, $M_i - M_i' \leqslant M - m$ and $M_i - M_i'' \leqslant M - m$. Consequently, $S - S' \leqslant (M - m)(\Delta x_i' + \Delta x_i'') = (M - m)\Delta x_i$. Since $\Delta x_i \leqslant \Delta$, we have $S - S' \leqslant (M - m)\Delta$. This inequality coincides with the first inequality given in the formulation of property 5°, for $p = 1$. The proof for the lower sums can be carried out in the same way.

6°. ***Darboux's Lemma.*** *Darboux's upper and lower integrals* $\bar{I}$ *and* I *of the function* $f(x)$ *with respect to the closed interval* $[a, b]$ *are the limits*** *of the lower and upper sums, respectively, for* $\Delta \to 0$.

Proof. Let us prove, for instance, that $\lim\limits_{\Delta \to 0} S = \bar{I}$. For the case $M = m$, that is for the case $f(x) = c = \text{const}$, the lemma is obvious inasmuch as $S = \bar{I} = I = s$. Therefore, we shall assume $M > m$. Since $\bar{I}$ is the infimum of the set of the upper sums, there is a partition T^* of the interval $[a, b]$, for any given $\varepsilon > 0$, such that the upper sum S^* of that partition differs from $\bar{I}$ by less than $\frac{\varepsilon}{2}$:

$$S^* - \bar{I} < \frac{\varepsilon}{2}. \tag{10.1}$$

Let us denote by p the number of points of partition T^* lying strictly in the interior of the closed interval $[a, b]$. Assume T to be any partition of the interval $[a, b]$ the maximum length Δ of whose subintervals obeys the condition

$$\Delta < \delta = \frac{\varepsilon}{2(M - m)p} \tag{10.2}$$

and S is the upper sum of that partition. Let us add to this partition the interior points of partition T^*. The result will be partition T' whose upper sum S', by virtue of property 5° and condition (10.2) for Δ, satisfies the inequality

$$0 \leqslant S - S' \leqslant (M - m)\, p\Delta < \frac{\varepsilon}{2}. \tag{10.3}$$

On the other hand, this partition T' may be considered as a partition obtained by adding the interior points of partition T to T^*. There-

---

* Somewhat earlier, while proving property 2°, we pointed out that the supremum of a function on a portion of a closed interval does not exceed its supremum on the entire interval. It should also be pointed out that the infimum of a function on the entire closed interval does not exceed its supremum on any portion of that interval.

** The concept of the limit of the upper and lower sums can be determined by a complete analogy with the concept of the limit of integral sums, namely, the number $\bar{I}$ is called the limit of the upper sums S for $\Delta \to 0$ if for any positive number ε there is a positive number δ such that for $\Delta < \delta$ the inequality $|S - \bar{I}| < \varepsilon$ holds true.

fore, by virtue of property 2°, we have

$$\overline{I} \leqslant S' \leqslant S^*.$$

Hence it follows that $0 \leqslant S' - \overline{I} \leqslant S^* - \overline{I}$, i. e. in accordance with inequality (10.1) we have

$$0 \leqslant S' - \overline{I} < \frac{\varepsilon}{2}.$$

Adding this inequality to inequality (10.3), we obtain

$$0 \leqslant S - \overline{I} < \varepsilon. \tag{10.4}$$

We have thus established that for any given positive ε there is a positive δ $\left(\text{we can put } \delta = \frac{\varepsilon}{2(M-m)p}, \text{ for instance}\right)$ such that the upper sums S of partition T of the interval $[a, b]$, for which the maximum length Δ of the subintervals is less than δ (see (10.2)), satisfy inequality (10.4). But this means that Darboux's upper integral $\overline{I}$ is the limit of the upper sums. The proof for the lower sums can be carried out by analogy. We have proved Darboux's lemma.

10.3. A NECESSARY AND SUFFICIENT CONDITION OF INTEGRABILITY

The properties of the upper and lower sums we have established above enable us to formulate, in a rather simple form, the necessary and sufficient condition of integrability of a function. To wit, the following *fundamental* theorem holds true.

Theorem 10.1. *For the function $f(x)$ bounded on the closed interval $[a, b]$ to be integrable on that interval, it is necessary and sufficient that for any positive ε there should be a partition T of the interval $[a, b]$ for which*

$$S - s \leqslant \varepsilon.$$

Proof. (1) **Necessity.** Let the function $f(x)$ be integrable on the closed interval $[a, b]$. We shall denote by I the limit of the integral sums of that function. By the definition of the limit of integral sums, there can be indicated $\delta > 0$, for any positive ε, such that for any partition T satisfying the condition $\Delta < \delta$ the inequality

$$|I\{x_i, \xi_i\} - I| < \frac{\varepsilon}{4} \tag{10.5}$$

holds true irrespective of the choice of points ξ_i on the subintervals of that partition.

Let us fix one of such partitions T. In accordance with property 1° (see 10.2.2), we can choose two integral sums or a given partition T (in other words, there can be points ξ_i' and ξ_i'' in each subin-

terval $[x_{i-1}, x_i]$) such that

$$S - I\{x_i, \xi_i'\} \leqslant \frac{\varepsilon}{4}, \quad I\{x_i, \xi_i''\} - s \leqslant \frac{\varepsilon}{4}.$$

Pay attention to the fact that both integral sums $I\{x_i, \xi_i'\}$ and $I\{x, \xi_i''\}$ satisfy inequality (10.5). The relation

$$S - s = (S - I\{x_i, \xi_i'\}) + (I\{x_i, \xi_i'\} - I) +$$
$$(I - I\{x_i, \xi_i''\}) + (I\{x_i, \xi_i''\} - s),$$

inequality (10.5) and the inequalities $S - I\{x_i, \xi_i'\} \leqslant \frac{\varepsilon}{4}$, $I\{x_i\xi_i''\} - s \leqslant \frac{\varepsilon}{4}$ yield

$$S - s < \varepsilon.$$

Thus we have proved the necessity of the conditions of the theorem.

(2) **Sufficiency.** Since inequalities $s \leqslant \underline{I} \leqslant \overline{I} \leqslant S$ are valid for any partition T and since for any positive ε, in accordance with the hypothesis, we can choose a partition such that $S - s \leqslant \varepsilon$, it follows that $0 \leqslant \overline{I} - \underline{I} \leqslant \varepsilon$. Since ε is arbitrary, we obtain equality $\underline{I} = \overline{I}$.

We shall denote the common value of the numbers $\underline{I}$ and $\overline{I}$ by I and prove that the number I is the limit of the integral sums of the function $f(x)$. Indeed, by virtue of Darboux's lemma (see 10.2.2), this number I is the common limit of the upper and lower sums for $\Delta \to 0$. Therefore, for any ε, we can choose a number δ such that for $\Delta < \delta$ the inequalities $I - s < \frac{\varepsilon}{2}$ and $S - I < \frac{\varepsilon}{2}$ hold true, that is $S - s < \varepsilon$ for $\Delta < \delta$, with $s \leqslant I \leqslant S$. Any integral sum $I\{x_i, \xi_i\}$ of the given partition T is contained between the upper sum and the lower sum $s \leqslant I\{x_i, \xi_i\} \leqslant S$. Thus, for $\Delta < \delta$, both values I and $I\{x_i, \xi_i\}$ are contained between the numbers S and s whose difference is less than ε. Hence it follows that for $\Delta < \delta$

$$|I\{x_i, \xi_i\} - I| < \varepsilon.$$

Consequently we infer that the number I is the limit of the integral sums. The theorem has been proved.

In what follows we shall need a somewhat different form of the notation for the necessary and sufficient condition of integrability. Assume that M_i and m_i are the supremum and the infimum of the values of the function $f(x)$ on the closed interval $[x_{i-1}, x_i]$. The number

$$\omega_i = M_i - m_i$$

is called the *oscillation of the function* $f(x)$ on the interval $[x_{i-1}, x_i]$. It should be pointed out that since $M_i \geqslant m_i$, the oscillation ω_i is a nonnegative number. Let us now write the difference $S - s$ in the following form:

$$S - s = \sum_{i=1}^{n} M_i \Delta x_i - \sum_{i=1}^{n} m_i \Delta x_i =$$

$$= \sum_{i=1}^{n} (M_i - m_i) \Delta x_i = \sum_{i=1}^{n} \omega_i \Delta x_i.$$

Since $\omega_i \geqslant 0$ and $\Delta x_i > 0$, each summand in the last sum is nonnegative.

A necessary and sufficient condition of integrability of a function can be written as follows:

For the function $f(x)$ *to be integrable on the closed interval* $[a, b]$, *it is necessary and suffiicient that there should be a partition* T *of the interval* $[a, b]$, *for any* $\varepsilon > 0$, *such that*

$$\sum_{i=1}^{n} \omega_i \Delta x_i \leqslant \varepsilon.$$

10.4. SOME CLASSES OF INTEGRABLE FUNCTIONS

In this section we shall prove integrability of functions continuous on a closed interval, of some discontinuous functions and of monotonic functions. To prove integrability of continuous functions, we shall need an important property of functions continuous on a closed interval which we shall now establish.

10.4.1. The property of uniform continuity of a function.

Definition. *The function* $f(x)$ *is said to be* **uniformly continuous** *on the set* $\{x\}$* *if for any positive number* ε *there is a positive number* δ, *dependent only on* ε, *such that for any two points* x' *and* x'' *of the set* $\{x\}$, *satisfying the condition* $|x'' - x'| < \delta$, *the inequality* $|f(x'') - f(x')| < \varepsilon$ *holds true.*

Remark. The main point in this definition is that for any $\varepsilon > 0$ there is a number $\delta > 0$ *guaranteeing* the validity of the inequality $|f(x'') - f(x')| < \varepsilon$ *for* x' *and* x'' *from the set* $\{x\}$ *all at once under the only condition that* $|x'' - x'| < \delta$.

Let us consider the following examples in order to elucidate the property of uniform continuity:

(1) The function $f(x) = \sqrt{x}$ *is uniformly continuous* on the half-line $x \geqslant 1$. Indeed, in accordance with Lagrange's theorem, we have

* In this case we assume that the set $\{x\}$ is dense in itself (see the end of 2.3).

for any $x' \geqslant 1$ and $x'' \geqslant 1$

$$|f(x'')-f(x')| = |f'(\xi)|\,|x''-x'| = \frac{1}{2\sqrt{\xi}}|x''-x'| < \frac{1}{2}|x''-x'|$$

(the last inequality follows from the fact that ξ is contained between x' and x'' and, therefore, $\xi > 1$). Consequently, if we proceed from a given positive ε and choose any δ satisfying the condition $0 < \delta \leqslant 2\varepsilon$, then for $|x''-x'| < \delta$ there holds an inequality $|f(x'')-f(x')| < \varepsilon$, that is on the set $x \geqslant 1$ the function $f(x) = \sqrt{x}$ is uniformly continuous.

(2) The function $f(x) = x^2$ *is not uniformly continuous* on the set $x \geqslant 1$. It suffices to prove that for some $\varepsilon > 0$ there is no $\delta > 0$ that would guarantee the validity of the inequality $|f(x'')-f(x')| < \varepsilon$ for all $x' \geqslant 1$ and $x'' \geqslant 1$ under the only condition that $|x''-x'| < \delta$. We shall prove that, in fact, even for any $\varepsilon > 0$ there is no δ indicated above. We shall fix $\varepsilon > 0$ and consider any positive δ. We shall choose $x' > \frac{\varepsilon}{\delta}$, $x'' = x' + \frac{\delta}{2}$. Then $|x''-x'| = \frac{\delta}{2} < \delta$. Making use of Lagrange's theorem, we obtain

$$|f(x'') - f(x')| = 2\xi\,|x'' - x'| = \xi\delta.$$

Since ξ is contained between x' and x'', we have $\xi > \frac{\varepsilon}{\delta}$ and, therefore, the last equality yields the inequality

$$|f(x'') - f(x')| > \varepsilon$$

although $|x'' - x'| < \delta$. Thus, the function $f(x) = x^2$ is not uniformly continuous on the set $x \geqslant 1$.

(3) The function $f(x) = \sin\frac{1}{x}$ *is not uniformly continuous* on the interval $(0, 1)$. We shall prove that for any ε satisfying the conditions $0 < \varepsilon < 2$, there is no $\delta > 0$ guaranteeing the validity of the inequality $|f(x'')-f(x')| < \varepsilon < 2$ for all x' and x'' belonging to the interval $(0, 1)$ under the only condition that $|x''-x'| < \delta$. To make sure of this fact, it is sufficient to put $x' = \frac{2}{(4k+3)\pi}$ and $x'' = \frac{2}{(4k+1)\pi}$ and to choose k, for any $\delta > 0$, so large that $|x''-x'| < \delta$. For the indicated points x' and x'' at any k, the difference

$$|f(x'')-f(x')| = \left|\sin\frac{1}{x''} - \sin\frac{1}{x'}\right| = 2 > \varepsilon.$$

Let us prove the following *fundamental* theorem.

Theorem 10.2 (theorem on uniform continuity). *The function $f(x)$ continuous on the closed interval $[a, b]$ is uniformly continuous on that interval.*

Proof. Let us assume that the function $f(x)$ continuous on the closed interval $[a, b]$ is not uniformly continuous on that interval. Then, for *some* $\varepsilon > 0$, the conditions formulated in the definition of uniform continuity are not satisfied. This means that for the indicated $\varepsilon > 0$ and *any* positive number δ belonging to the interval $[a, b]$ there are points x' and x'' such that $|x'' - x'| < \delta$ but $|f(x'') - f(x')| \geqslant \varepsilon$. Therefore, for every $\delta = \frac{1}{n}$, $n = 1, 2, \ldots$, there are points x'_n and x''_n from the interval $[a, b]$ such that $|x''_n - x'_n| < \frac{1}{n}$, but $|f(x''_n) - f(x'_n)| \geqslant \varepsilon$. Since $\{x'_n\}$ is a sequence of points belonging to the closed interval $[a, b]$, from this sequence we can separate, in accordance with the Bolzano-Weierstrass theorem, a subsequence $\{x'_{k_n}\}$ converging to some point c of that interval (see Remark 2 in 3.4.4). The subsequence $\{x''_{k_n}\}$ of the sequence $\{x''_n\}$ evidently converges to c as well. Since the function $f(x)$ is continuous at point c, the limits of the sequences $\{f(x'_{k_n})\}$ and $\{f(x''_{k_n})\}$ are equal to $f(c)$ and, therefore, the sequence $\{f(x''_{k_n}) - f(x'_{k_n})\}$ is infinitesimal. But it cannot be so since all the terms $f(x''_{k_n}) - f(x'_{k_n})$ of the indicated sequence satisfy the inequality $|f(x''_{k_n}) - f(x'_{k_n})| \geqslant \varepsilon$. Thus, the assumption that the function continuous on the interval $[a, b]$ is not uniformly continuous leads to a contradiction. We have proved the theorem.

Corollary. *Suppose the function $f(x)$ is continuous on the closed interval $[a, b]$. Then, for any positive number ε, there can be indicated a positive δ such that on each subinterval $[c, d]$ belonging to the interval $[a, b]$, whose length $d - c$ is less than δ, the oscillation ω* of the function $f(x)$ is less than ε.*

Proof. By virtue of the theorem we have just proved, the function $f(x)$ continuous on the closed interval $[a, b]$ is uniformly continuous on that interval. Therefore, for any $\varepsilon > 0$, we can choose a positive δ such that for any x' and x'' belonging to the interval $[a, b]$ and satisfying the condition $|x'' - x'| < \delta$ the inequality $|f(x'') - f(x')| < \varepsilon$ holds true. We shall prove now that on each subinterval $[c, d]$, belonging to the closed interval $[a, b]$, whose length $d - c$ is less than the indicated δ, the oscillation ω of the function $f(x)$ is less than ε. Indeed, since the function $f(x)$ is continuous on the interval $[c, d]$, there are points x' and x'' on that interval such that $f(x') = m$ and $f(x'') = M$, where m and M are the greatest lower bound

* Recall that the oscillation ω of the function $f(x)$ on the closed interval $[c, d]$ is the difference $M - m$ between the supremum and the infimum of the function $f(x)$ on that interval.

and the least upper bound of the function $f(x)$ on the interval $[c, d]$ (see Theorem 8.8). Since $|x'' - x'| < \delta$ (for the length of the interval $[c, d]$ is less than δ), we have $|f(x'') - f(x')| < \varepsilon$. But $f(x'') - f(x') = M - m = \omega$. Consequently, $\omega < \varepsilon$.

Remark. The set $\{x\}$ of the points of the number line is said to be *closed* if it contains all its limit points*. The following assertion is valid. *A function $f(x)$ continuous on the bounded closed** set $\{x\}$ is uniformly continuous on that set.* The proof of this assertion is analogous to the proof of Theorem 10.2.

10.4.2. Heine-Borel lemma. Another proof of the theorem on uniform continuity. Point x of the set $\{x\}$ is called an *interior point* of that set, if it belongs to a certain interval whose all points belong to the set $\{x\}$. The set $\{x\}$ is said to be *open* if all the points of that set are interior points.

We shall say that the given set $\{x\}$ *is covered by a system* $\sum$ *of open sets*** if every point x of that set belongs to at least one set of the system $\sum$. Let us prove the following lemma.

Heine-Borel** **lemma.** *If a closed interval $[a, b]$ is covered by an infinite system* $\sum$ *of open sets, then a finite subsystem* $\sum'$ *of sets which also covers the interval $[a, b]$ can be isolated from that system.*

Proof.*** Let $\{x\}$ be as et of points of the interval $[a, b]$ such that if x belongs to this set, the interval $[a, x]$ is covered by some finite subsystem $\sum'$ of the sets of the system $\sum$. Let us prove that the set $\{x\}$ coincides with the interval $[a, b]$. Since point a is covered by a set of the system $\sum$ and this set is open, it also covers some closed interval $[a, x]$ whose all points belong to the set $\{x\}$ in accordance with what was said above. The set $\{x\}$ is, evidently, bounded. Suppose $\bar{x} = \sup\{x\}$. Let us verify that $\bar{x}$ belongs to the set $\{x\}$ and that $\bar{x} = b$. Indeed, $\bar{x}$ is covered by a set of the system $\sum$ and, consequently, all the points of some interval $(\bar{x} - \varepsilon, \bar{x} + \varepsilon)$ are covered by the same set. Since $\bar{x} = \sup\{x\}$, there are points of the set $\{x\}$ arbitrarily close to $\bar{x}$, and, therefore, we can find a point x' of that set belonging to the interval $(\bar{x} - \varepsilon, \bar{x} + \varepsilon)$. It follows from the definition of the set $\{x\}$ that the closed interval $[a, x']$ is covered by some finite subsystem $\sum'$ of the sets of the system $\sum$. Adding to $\sum'$ the set covering points $\bar{x}$, we shall obtain a finite subsystem $\sum''$ of the sets of the system $\sum$ which covers the interval $[a, \bar{x}]$. Hence we infer that $\bar{x}$ belongs to $\{x\}$. If we assumed that $\bar{x} < b$, the subsystem $\sum''$ would cover all the points of a certain interval $[a, x'']$, where $\bar{x} < x'' < \bar{x} + \varepsilon$ and, therefore, point x'' would belong to the set $\{x\}$. But that cannot be so since $\bar{x}$ is the least upper bound of the set $\{x\}$. Thus, the set $\{x\}$ coincides with the interval $[a, b]$. We have proved the lemma.

---

* The definition of the limit point of a set was given in 3.4.6.

** You can find the definition of a bounded set in 3.4.6.

*** If the set $\{x\}$ consists of one point and the system Σ contains only one open set, then we shall say that this set covers the indicated point.

**** Eduard Heine (1821-1881), a German mathematician, Emile Borel (1871-1956), a French mathematician.

***** This proof of the Heine-Borel lemma belongs to the French mathematician Henri Lebesgue (1875-1941). Note that Lebesgue pointed out and substantiated a more general approach to the problem of integration than that given in the present chapter. The corresponding concept of the integral is called the Lebesgue integral.

Remark. The Heine-Borel lemma can be generalized as follows. *If a closed* bounded set* $\{x\}$ *is covered by an infinite system* $\sum$ *of open sets, then a finite subsystem* $\overline{\sum}$ *of sets can be isolated from this system which also covers the set* $\{x\}$. We shall now give another proof of Theorem 10.2.

Proof of the theorem on uniform continuity. Let us extend $f(x)$ to the whole straight line putting it equal to $f(b)$ for $x > b$ and to $f(a)$ for $x < a$. Since $f(x)$ is continuous at each point of the closed interval $[a, b]$, we can choose, for any point x of that interval and any given $\varepsilon > 0$, a number $\delta' > 0$ depending, in a general case, on x, such that for all the points x' satisfying the condition $|x' - x| < \delta'$ the inequality $|f(x') - f(x)| < \frac{\varepsilon}{2}$ holds true. Thus we see, that the interval $[a, b]$ is covered by an infinite system $\sum$ of intervals $\left(x - \frac{\delta'}{2}, x + \frac{\delta'}{2}\right)$** from which we can isolate, by virtue of the Heine-Borel lemma, a finite subsystem $\overline{\sum}$ of intervals which also covers the interval $[a, b]$. Let us designate by δ the minimum value $\frac{\delta'}{2}$ for this finite subsystem $\overline{\sum}$ of intervals. Assume now that x' and x'' are any points belonging to the interval $[a, b]$ and satisfying the condition $|x'' - x'| < \delta$, and x is the centre of the interval $\left(x - \frac{\delta'}{2}, x + \frac{\delta'}{2}\right)$, $\delta \leqslant \frac{\delta'}{2}$, of the system $\overline{\sum}$ which covers point x'. Since $|x' - x| < \frac{\delta'}{2}$ and $|x'' - x| < \delta'$, we have $|f(x') - f(x)| < \frac{\varepsilon}{2}$ and $|f(x'') - f(x)| < \frac{\varepsilon}{2}$ and, therefore,

$$|f(x'') - f(x')| \leqslant |f(x'') - f(x)| + |f(x') - f(x)| < \\ < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Thus, for any given positive ε we have indicated a positive δ such that for any points x' and x'' of the closed interval $[a, b]$ satisfying the condition $|x'' - x'| < \delta$ the inequality $|f(x'') - f(x')| < \varepsilon$ holds true. Consequently, the function $f(x)$ is uniformly continuous on the interval $[a, b]$. And that is the proof of the theorem.

10.4.3. Integrability of continuous functions. We shall now prove the following *fundamental* theorem.

Theorem 10.3. *A function* $f(x)$ *continuous on the closed interval* $[a, b]$ *is integrable on that interval.*

Proof. Assume that we are given an arbitrary positive ε. By virtue of uniform continuity of the function $f(x)$ on the interval $[a, b]$, for the positive number $\frac{\varepsilon}{b-a}$ there is a number $\delta > 0$ such that upon partitioning T of the interval $[a, b]$ into subintervals $[x_{i-1}, x_i]$ whose lengths Δx_i are less than δ, the oscilla-

* See the remark in the previous subsection.

** We take the intervals $\left(x - \frac{\delta'}{2}, x + \frac{\delta'}{2}\right)$ instead of $(x - \delta', x + \delta')$ to make the argument more convenient.

tion ω_i of the function $f(x)$ on each of these subintervals is less than $\frac{\varepsilon}{b-a}$ (see the corollary of Theorem 10.2). Therefore, for such partitions T

$$S-s=\sum_{i=1}^{n}\omega_i\Delta x_i<\frac{\varepsilon}{b-a}\sum_{i=1}^{n}\Delta x_i=\varepsilon.$$

Consequently, for the function $f(x)$ continuous on the interval $[a, b]$ the sufficient conditions for integrability are satisfied.

10.4.4. Integrability of some discontinuous functions. We shall say that point x is covered by an interval if this point belongs to the indicated interval. Let us prove the following theorem.

Theorem 10.4. *If the function $f(x)$ is defined and bounded on the interval $[a, b]$ and if for any positive number ε we can indicate a finite number of intervals covering all the points of discontinuity of that function and having the total sum of lengths less than ε, then $f(x)$ is integrable on the interval $[a, b]$.*

Proof. Assume that we are given any $\varepsilon>0$. Let us cover the points of discontinuity of the function $f(x)$ by a finite number of intervals the sum of whose lengths is less than $\frac{\varepsilon}{2(M-m)}$, where M and m are the supremum and the infimum of $f(x)$ on the interval $[a, b]$ (the case $M=m$ can be excluded from the discussion since then $f(x)\equiv c\equiv\text{const}$). The points of the closed interval not belonging to the indicated intervals constitute a set consisting of a finite number of nonintersecting closed intervals. On each of them the function $f(x)$ is continuous and, therefore, uniformly continuous. Let us divide each of these closed intervals so that the oscillation ω_i of the function $f(x)$ on any subinterval is less than $\frac{\varepsilon}{2(b-a)}$. Combining these subintervals and the intervals covering the points of discontinuity of the function $f(x)$, we obtain partition T of the whole closed interval $[a, b]$. For this partition, the summands of the sum $\sum_{i=1}^{n}\omega_i\Delta x_i$ (equal to $S-s$) are divided into two groups $\sum'\omega_i\Delta x_i$ and $\sum''\omega_i\Delta x_i$, with the first group containing all the summands corresponding to the parts of partition T formed by the intervals covering the points of discontinuity, and with the second group containing the rest of the summands. Since the oscillations $\omega_i=M_i-m_i$ for the summands of the first group satisfy the inequality $\omega_i\leqslant M-m$, we have

$$\sum{}'\omega_i\Delta x_i\leqslant(M-m)\sum{}'\Delta x_i<(M-m)\frac{\varepsilon}{2(M-m)}=\frac{\varepsilon}{2}.$$

For the summands of the second group $\omega_i < \frac{\varepsilon}{2(b-a)}$. Therefore,

$$\sum'' \omega_i \Delta x_i < \frac{\varepsilon}{2(b-a)} \sum'' \Delta x_i \leqslant \frac{\varepsilon}{2(b-a)} (b-a) = \frac{\varepsilon}{2}.$$

Thus we have

$$S - s = \sum_{i=1}^{n} \omega_i \Delta x_i = \sum' \omega_i \Delta x_i + \sum'' \omega_i \Delta x_i < \varepsilon.$$

Thus sufficient conditions of integrability are satisfied for the function $f(x)$ indicated in the hypothesis. The theorem is proved.

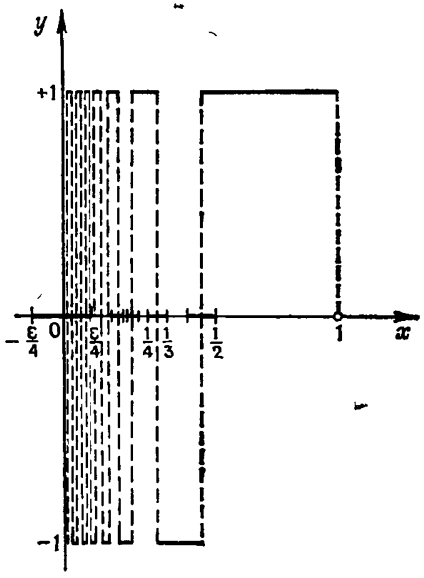


Fig. 10.4

Corollary. The function $f(x)$ bounded on the closed interval $[a, b]$ which has only a finite number of points of discontinuity is integrable on that interval. In particular, a function piecewise continuous on a given closed interval is integrable on that interval.*

Remark. It is evident that if the function $f(x)$ is integrable on the closed interval $[a, b]$ and the function $g(x)$ differs from $f(x)$ only at a finite number of points, the function $g(x)$ is also integrable on the interval $[a, b]$, with $\int_a^b f(x)\,dx = \int_a^b g(x)\,dx$.

Let us consider an example of an integrable function having an *infinite* number of points of discontinuity. Suppose a function $f(x)$ is given on the closed interval $[0, 1]$ (Fig. 10.4):

$$f(x) = \begin{cases} 1 \text{ on semi-open intervals } \left(\frac{1}{2n}, \frac{1}{2n-1}\right], n = 1, 2, \ldots, \\ -1 \text{ on semi-open intervals } \left(\frac{1}{2n+1}, \frac{1}{2n}\right], n = 1, 2, \ldots, \\ 0 \text{ at point } x = 0. \end{cases}$$

* If p is the number of points of discontinuity, it is sufficient to cover each point of discontinuity by an interval of length $\varepsilon/2p$.

The indicated function possesses discontinuities of the 1st kind at *all* the points $x_n=\frac{1}{n}$, $n=2, 3, \ldots$. Let us fix any $\varepsilon>0$. We shall cover point $x=0$ (in any neighbourhood of this point there are an infinite number of points of discontinuity of the function) by an interval $\left(-\frac{\varepsilon}{4}, \frac{\varepsilon}{4}\right)$. Outside this interval there are only a finite number p* of points of discontinuity of the function each of which we shall cover by an interval of length less than $\frac{\varepsilon}{2p}$. The sum of the lengths of the intervals covering all the points of discontinuity of the function being considered is less than $\frac{\varepsilon}{2}+p\frac{\varepsilon}{2p}=\varepsilon$. Consequently, the function $f(x)$ is integrable on the closed interval $[0, 1]$.

10.4.5. Integrability of monotonic bounded functions.

*Theorem 10.5. The function $f(x)$ monotonicon the closed interval $[a, b]$ is integrable on that interval***.*

Proof. For the sake of definiteness, we shall prove the theorem for a function $f(x)$ nondecreasing on that interval $[a, b]$. We shall take an arbitrary positive number ε and partition the interval $[a, b]$ into equal parts whose lengths are less than $\frac{\varepsilon}{f(b)-f(a)}$ (the case $f(a)=f(b)$ can be excluded since then $f(x)=$ $=$const). Let us evaluate for that partition the difference $S-s=\sum_{i=1}^{n}\omega_i\Delta x_i$. We obtain

$$S-s=\sum_{i=1}^{n}\omega_i\Delta x_i<\frac{\varepsilon}{f(b)-f(a)}\sum_{i=1}^{n}\omega_i.$$

But for a nondecreasing function $\sum_{i=1}^{n}\omega_i=f(b)-f(a)$, therefore, $S-s<\varepsilon$. We have proved the theorem.

10.5. THE MAIN PROPERTIES OF THE DEFINITE INTEGRAL

Let us prove the validity of the following properties of the definite integral:

1°. We shall assume that

$$\int_a^a f(x)\,dx=0. \tag{10.6}$$

* This number p is naturally dependent on ε.

** It should be pointed out that if a function is monotonic on the closed interval $[a, b]$ then its values are contained between $f(a)$ and $f(b)$. Therefore, the monotonic function defined on the interval $[a, b]$ is bounded on that interval.

It might be well to point out that formula (10.6) should be considered as a sort of agreement. It should be taken as a natural extension of the concept of the definite integral to a closed interval of zero length.

2°. We shall assume that for $a < b$

$$\int_b^a f(x)\,dx = -\int_a^b f(x)\,dx. \tag{10.7}$$

This formula must also be treated as an agreement. It is a natural generalization of the concept of the integral to the case where the closed interval $[a, b]$, $a < b$, is run through from b to a (in that case all the differences $\Delta x_i = x_i - x_{i-1}$ in the integral sum have a negative sign).

3°. Assume that the functions $f(x)$ and $g(x)$ are integrable on the closed interval $[a, b]$. Then the functions $f(x) + g(x)$, $f(x) - g(x)$ and $f(x)g(x)$ are also integrable on that interval, with

$$\int_a^b [f(x) \pm g(x)]\,dx = \int_a^b f(x)\,dx \pm \int_a^b g(x)\,dx. \tag{10.8}$$

We shall first prove integrability of the function $f(x) \pm g(x)$ and the validity of formula (10.8). With any partition of the interval $[a, b]$ and any choice of points ξ_i the relation

$$\sum_{i=1}^{n} [f(\xi_i) \pm g(\xi_i)]\,\Delta x_i = \sum_{i=1}^{n} f(\xi_i)\,\Delta x_i \pm \sum_{i=1}^{n} g(\xi_i)\,\Delta x_i$$

is valid for integral sums and, therefore, it follows from the existence of the limit of the right-hand side that the limit of the left-hand side also exists. Consequently, the function $f(x) \pm g(x)$ is integrable and formula (10.8) is valid.

Let us now prove that the product of integrable functions is an integrable function. Since the functions $f(x)$ and $g(x)$ are integrable on the closed interval $[a, b]$, they are bounded on that interval (see the statement in 10.1.1), so that $|f(x)| \leqslant A$ and $|g(x)| \leqslant B$. Let us consider any given partition T of the interval $[a, b]$. Assume that x' and x'' are arbitrary points of the subinterval $[x_{i-1}, x_i]$. We have an identity

$$f(x'')\,g(x'') - f(x')\,g(x') = [f(x'') - f(x')]\,g(x'') + [g(x'') - g(x')]\,f(x').$$

Since

$$|f(x'')\,g(x'') - f(x')\,g(x')| \leqslant \omega_i,$$

$$|f(x'') - f(x')| \leqslant \bar{\omega}_i, \quad |g(x'') - g(x')| \leqslant \bar{\bar{\omega}}_i,$$

where ω_i, $\bar{\omega}_i$, $\bar{\bar{\omega}}_i$ are, respectively, the oscillations of the functions $f(x)\,g(x)$, $f(x)$, $g(x)$ on the interval $[x_{i-1}, x_i]$, we have, in accordance with the indicated identity*,

$$\omega_i \leqslant B\bar{\omega}_i + A\bar{\bar{\omega}}_i.$$

Therefore,

$$\sum_{i=1}^{n} \omega_i \Delta x_i \leqslant B \sum_{i=1}^{n} \bar{\omega}_i \Delta x_i + A \sum_{i=1}^{n} \bar{\bar{\omega}}_i \Delta x_i.$$

Since the functions $f(x)$ and $g(x)$ are integrable on the closed interval $[a, b]$, we can choose partition T of this interval, for any given positive ε, such that $\sum_{i=1}^{n} \bar{\omega}_i \Delta x_i < \frac{\varepsilon}{2B}$ and $\sum_{i=1}^{n} \bar{\bar{\omega}}_i \Delta x_i < \frac{\varepsilon}{2A}$. Consequently, we have for this partition

$$S - s = \sum_{i=1}^{n} \omega_i \Delta x_i < B\frac{\varepsilon}{2B} + A\frac{\varepsilon}{2A} = \varepsilon.$$

Therefore, a product of integrable functions is an integrable function.

4°. If the function $f(x)$ is integrable on the closed interval $[a, b]$, then the function $cf(x)$ (c = const) is integrable on that interval, with

$$\int_a^b cf(x)\,dx = c\int_a^b f(x)\,dx. \tag{10.9}$$

Indeed, the integral sums of the functions $f(x)$ and $cf(x)$ differ only by the constant factor c. Therefore, the function $cf(x)$ is integrable and formula (10.9) is valid.

5°. Assume that the function $f(x)$ is integrable on the closed interval $[a, b]$. Then this function is integrable on any interval $[c, d]$ contained in the interval $[a, b]$.

Since the function $f(x)$ is integrable on the closed interval $[a, b]$, there exists, for any $\varepsilon > 0$, a partition T of the interval $[a, b]$ such that $S - s < \varepsilon$ (see Theorem 10.1). Let us add points c and d to the points of partition T. By virtue of property 2° of the upper and lower sums (see 10.2.2), the inequality $S - s < \varepsilon$ s obviously valid for partition T^*. This partition of the interval $[a, b]$ brings about partition $\bar{T}$ of the interval $[c, d]$. If $\bar{S}$ and $\bar{s}$ are the upper and

* In this identity the points x' and x'' can be chosen such that the left-hand side will be arbitrarily close to ω_i.

the lower sum of partition $\overline{T}$, then $\overline{S} - \overline{s} \leqslant S - s$, since every nonnegative summand $\omega_i \Delta x_i$ in the expression $\overline{S} - \overline{s} = \sum \omega_i \Delta x_i$ will also be a summand in the expression for $S - s$. Consequently, $\overline{S} - \overline{s} < \varepsilon$, and, therefore, the function $f(x)$ is integrable on the closed interval $[c, d]$.

6°. Assume that the function $f(x)$ is integrable on the closed intervals $[a, c]$ and $[c, b]$. Then this function is integrable on the interval $[a, b]$, with

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx. \tag{10.10}$$

We shall first consider the case where $a < c < b$. Since the function $f(x)$ is integrable on the intervals $[a, c]$ and $[c, b]$, there are partitions of those intervals such that the difference $S - s$ for each of them is less than $\frac{\varepsilon}{2}$. Combining these partitions, we obtain a partition of the interval $[a, b]$ for which the difference $S - s$ is less than ε. Consequently, the function $f(x)$ is integrable on $[a, b]$. We shall include point c into the number of points of division of the interval $[a, b]$ at its every partitioning. Then, the integral sum for $f(x)$ on $[a, b]$ is equal to the sum of the integral sums for that function on $[a, c]$ and $[c, b]$. In the limit, we shall obtain formula (10.10).

If point c lies outside of the interval $[a, b]$, then the interval $[a, b]$ is a part of the interval $[a, c]$ (or $[c, b]$) and, therefore, by virtue of property 5°, the function $f(x)$ is integrable on the interval $[a, b]$. Let us consider the case $a < b < c$. Then we have

$$\int_a^b f(x)\,dx + \int_b^c f(x)\,dx = \int_a^c f(x)\,dx,$$

whence, making use of property 2° and formula (10.7), we shall again obtain relation (10.10). It is easy to verify the validity of this relation in the case $c < a < b$ as well.

10.6. EVALUATIONS OF INTEGRALS. MEAN-VALUE FORMULAS

10.6.1. Evaluations of integrals. In this section we shall obtain some evaluations of definite integrals whose integrand functions are subject to some or other conditions.

1°. *Suppose the function $f(x)$ integrable on the closed interval $[a, b]$ is nonnegative on that interval. Then we have*

$$\int_a^b f(x)\,dx \geqslant 0.$$

Indeed, every integral sum of such a function is nonnegative and, therefore, the limit $I=\int_a^b f(x)\,dx$ of the integral sums is also nonnegative *.

Remark 1. *If $f(x)$ is integrable on the closed interval $[a, b]$ and $f(x)\geqslant m$, then*

$$\int_a^b f(x)\,dx\geqslant m(b-a).$$

Indeed, the function $f(x)-m\geqslant 0$ and it is integrable on the closed interval $[a, b]$. Therefore, $\int_a^b [f(x)-m]\,dx\geqslant 0$. Hence we have

$$\int_a^b f(x)\,dx\geqslant\int_a^b m\,dx=m\int_a^b dx=m(b-a)$$

(see property 3° and the example in 10.1).

2°. *If the function $f(x)$ is continuous, nonnegative and is not identically zero on the closed interval $[a, b]$, then*

$$\int_a^b f(x)\,dx\geqslant c>0.$$

Indeed, since the function $f(x)$ is nonnegative and is not identically zero, there is a point ξ on the closed interval $[a, b]$ such that $f(\xi)=2k>0$. Then, in accordance with the theorem on the stability of sign of a continuous function, there can be found a closed interval $[p, q]$ containing point ξ in its interior, such that the values of the function $f(x)$ throughout the interval will not be less than $k>0$. Therefore, by virtue of the remark we have just made,

$$\int_p^q f(x)\,dx\geqslant k(q-p)>0.$$

In accordance with property 6° of definite integrals

$$\int_a^b f(x)\,dx=\int_a^p f(x)\,dx+\int_p^q f(x)\,dx+\int_q^b f(x)\,dx.$$

* Let us assume that the limit I of the integral sums is negative. Then, in accordance with the definition of the limit I, there can be found, for the number $\varepsilon=|I|$, an integral sum $I\{x_i, \xi_i\}$, for which $|I\{x_i, \xi_i\}-I|<|I|$. This inequality yields $I\{x_i, \xi_i\}<0$, and so we have just proved that every integral sum is nonnegative. Consequently, the limit I is nonnegative.

Therefore, since $f(x) \geqslant 0$ and $\int_p^q f(x)\,dx \geqslant c > 0$, where $c = k(q-p)$,

$$\int_a^b f(x)\,dx \geqslant c > 0.$$

3°. *If the functions $f(x)$ and $g(x)$ are integrable on the closed interval $[a, b]$ and $f(x) \geqslant g(x)$ everywhere on that interval, we have*

$$\int_a^b f(x)\,dx \geqslant \int_a^b g(x)\,dx.$$

Indeed, the function $f(x) - g(x) \geqslant 0$ and is integrable on the interval $[a, b]$. Hence follows the validity of the indicated evaluation, by virtue of property 1°.

Remark 2. *If the function $f(x)$ is integrable on the closed interval $[a, b]$, then the function $|f(x)|$ is also integrable on that interval, with*

$$\left|\int_a^b f(x)\,dx\right| \leqslant \int_a^b |f(x)|\,dx.$$

We shall first prove the integrability of the absolute value $|f(x)|$ of the integrable function $f(x)$. We shall designate by M_i and m_i the supremum and the infimum of $f(x)$ on the closed interval $[x_{i-1}, x_1]$, and by M_i' and m_i' the supremum and the infimum of $|f(x)|$ on the same interval. It is easy to verify that $M_i' - m_i' \leqslant M_i - m_i$ (it is sufficient to consider three possible cases: (1) M_i and m_i are nonnegative, (2) M_i and m_i are nonpositive, (3) $M_i > 0$, $m_i \leqslant 0$). From the inequality obtained it follows that $S' - s' \leqslant S - s$. Thus, if for some partition $S - s < \varepsilon$, then for that partition $S' - s' < \varepsilon$, that is, a sufficient condition of integrability for $|f(x)|$ is fulfilled*.

Let us now prove the evaluation we are interested in. Since $-|f(x)| \leqslant f(x) \leqslant |f(x)|$, we have $-\int_a^b |f(x)|\,dx \leqslant \int_a^b f(x)\,dx \leqslant$

---

* Integrability of the function $|f(x)|$ does not yield, in a general case, integrability of $f(x)$. For instance, $f(x) = \begin{cases} 1 \text{ for rational } x, \\ -1 \text{ for irrational } x, \end{cases}$ is not integrable on the closed interval $[0, 1]$ whereas $|f(x)| \equiv 1$ is a function integrable on that interval.

$\leqslant \int_a^b |f(x)|\,dx$, and this precisely means that $\left|\int_a^b f(x)\,dx\right| \leqslant \leqslant \int_a^b |f(x)|\,dx$.

4°. *Suppose the functions $f(x)$ and $g(x)$ are integrable on the closed interval $[a, b]$ and $g(x) \geqslant 0$. Then in case M and m are the supremum and the infimum of $f(x)$ on the interval $[a, b]$, we have*

$$m\int_a^b g(x)\,dx \leqslant \int_a^b f(x)\,g(x)\,dx \leqslant M\int_a^b g(x)\,dx. \tag{10.11}$$

The validity of (10.11) follows from the fact that inequalities $mg(x) \leqslant f(x)\,g(x) \leqslant Mg(x)$ hold true for all x belonging to the interval $[a, b]$ (see evaluation 3° from the present subsection and property 4° in 10.5).

Remark. 3. In Supplement 1 to the present chapter we shall obtain some important inequalities for sums and for definite integrals.

10.6.2. The first mean-value formula. Assume that the function *$f(x)$ is integrable on the closed interval $[a, b]$ and let m and M be the infimum and the supremum of $f(x)$ on the interval $[a, b]$. Then, there can be chosen a number μ, satisfying the inequalities $m \leqslant \mu \leqslant M$, such that*

$$\int_a^b f(x)\,dx = \mu(b-a). \tag{10.12}$$

Indeed, putting $g(x) = 1$ and taking into account that $\int_a^b 1 \cdot dx = b - a$ (see the example in 10.1.1), we shall obtain, from (10.11),

$$m(b-a) \leqslant \int_a^b f(x)\,dx \leqslant M(b-a).$$

Denoting by μ the number $\frac{1}{b-a}\int_a^b f(x)\,dx$, we shall have formula (10.12).

If the function $f(x)$ is *continuous* on the closed interval $[a, b]$, there exist points p and q of that interval such that $f(p) = m$ and $f(q) = M$ (see Theorem 8.8) and, therefore, by virtue of Theorem 8.6, there is a point ξ on the interval $[p, q]$ and hence, on $[a, b]$,

such that $f(\xi) = \mu$. In this case formula (10.12) assumes the form

$$\int_a^b f(x)\,dx = f(\xi)\,(b-a). \tag{10.13}$$

This formula is the *first mean-value formula.*

10.6.3. The first mean-value formula in a generalized form. Let us prove the following assertion. *Suppose the functions $f(x)$ and $g(x)$ are integrable on the closed interval $[a, b]$ and m and M are the infimum and the supremum of $f(x)$ on the interval $[a, b]$. Assume, in addition, that $g(x) \geqslant 0$ (or $g(x) \leqslant 0$) throughout the interval $[a, b]$. Then, there can be chosen a number μ, satisfying the inequalities $m \leqslant \mu \leqslant M$, such that*

$$\int_a^b f(x)\,g(x)\,dx = \mu \int_a^b g(x)\,dx. \tag{10.14}$$

In particular, if $f(x)$ is continuous on the closed interval $[a, b]$ there is a number ξ on that interval such that

$$\int_a^b f(x)\,g(x)\,dx = f(\xi) \int_a^b g(x)\,dx. \tag{10.15}$$

Formula (10.15) is called the *first mean-value formula in a generalized form.*

Let us prove the validity of formula (10.14). If $\int_a^b g(x)\,dx = 0$, then, by virtue of inequalities (10.11), $\int_a^b f(x)\,g(x)\,dx = 0$ and, therefore, we can take any number as μ. If now $\int_a^b g(x)\,dx > 0$, then, having divided all the parts of inequalities (10.11) by $\int_a^b g(x)\,dx$, we shall obtain

$$m \leqslant \frac{\int_a^b f(x)\,g(x)\,dx}{\int_a^b g(x)\,dx} \leqslant M.$$

Denoting by μ the number $\dfrac{\int_a^b f(x)\,g(x)\,dx}{\int_a^b g(x)\,dx}$, we shall obtain formula (10.14).

If $f(x)$ is continuous on the closed interval $[a, b]$, then, whatever the number μ contained between m and M, there is a point ξ on that interval such that $f(\xi) = \mu$, that is formula (10.14) passes into formula (10.15).

Remark 4. If the function $f(x)$ is not continuous, then formula (10.15), generally speaking, is not valid. Indeed, suppose, for instance, that $f(x) = \begin{cases} \frac{1}{2} & \text{for } 0 \leqslant x \leqslant \frac{1}{2}, \\ 1 & \text{for } \frac{1}{2} < x \leqslant 1, \end{cases}$ and $g(x) = \begin{cases} 1 & \text{for } 0 \leqslant x \leqslant \frac{1}{2}, \\ \frac{1}{2} & \text{for } \frac{1}{2} < x \leqslant 1. \end{cases}$ Then, as it is easy to verify the number μ in formula (10.14) is equal to $\frac{2}{3}$. Thus, for any ξ belonging to the interval $[0, 1]$ $f(\xi) \neq \mu$.

10.6.4. The second mean-value formula. The following assertion is valid. *If the function $g(x)$ is monotonic on the closed interval $[a, b]$ and $f(x)$ is integrable on that interval, then there is a number ξ on that interval such that*

$$\int_a^b f(x)\,g(x)\,dx = g(a)\int_a^{\xi} f(x)\,dx + g(b)\int_{\xi}^{b} f(x)\,dx. \qquad (10.16)$$

Formula (10.16) is the *second mean-value formula* or *Bonnet's* formula*. The assertion we have formulated will be proved in Supplement 2 to this chapter.

10.7. THE EXISTENCE OF AN ANTIDERIVATIVE OF A CONTINUOUS FUNCTION. BASIC RULES OF INTEGRATION

10.7.1. The existence of an antiderivative of a continuous function. Before proceeding to the proof of a theorem on the existence of an antiderivative for a continuous function, we shall introduce the concept of an *integral with a variable upper limit.*

Suppose the function $f(x)$ is integrable on any closed interval contained in the interval (a, b) and assume that c is some fixed point be-

* P. O. Bonnet (1819-1892), a French mathematician.

longing to that interval. Then, whatever the number x from the interval (a, b), the function $f(x)$ is integrable on the closed interval $[c, x]$. Therefore, the function

$$F(x)=\int_c^x f(t)\,dt^*$$

is defined on the interval (a, b), and this function is called the *integral with a variable upper limit.* Let us prove the following theorem.

Theorem 10.6. *Any function $f(x)$ continuous on the interval (a, b) possesses an antiderivative on that interval. One of the antiderivatives is the function*

$$F(x)=\int_c^x f(t)\,dt,$$

where c is any fixed point from the interval (a, b).

Proof. It is sufficient to prove that for any fixed x belonging to the interval (a, b) there exists a limiting value $\lim\limits_{\Delta x\to 0}\dfrac{F(x+\Delta x)-F(x)}{\Delta x}$, and this limiting value is equal to $f(x)$. By virtue of property 6° of definite integrals (see 10.5)**, we have

$$F(x+\Delta x)-F(x)=\int_c^{x+\Delta x} f(t)\,dt-\int_c^x f(t)\,dt=$$

$$=\int_c^x f(t)\,dt+\int_x^{x+\Delta x} f(t)\,dt-\int_c^x f(t)\,dt=\int_x^{x+\Delta x} f(t)\,dt.$$

We find by mean-value formula (10.13) that

$$F(x+\Delta x)-F(x)=\int_x^{x+\Delta x} f(t)\,dt=f(\xi)\,\Delta x,$$

where ξ is a number contained between the numbers x and $x+\Delta x$. Since the function $f(x)$ is continuous at point x, we have $f(\xi)\to f(x)$ for $\Delta x\to 0$. Therefore, the last formula yields

$$\lim_{\Delta x\to 0}\frac{F(x+\Delta x)-F(x)}{\Delta x}=\lim_{\Delta x\to 0} f(\xi)=f(x).$$

We have proved the theorem.

---

* We have denoted the variable of integration by the letter t since the letter x is used to denote the upper limit of integration.

** We take Δx small enough for $(x+\Delta x)$ to belong to (a, b).

Remark 1. The theorem on the existence of an antiderivative of a function continuous *on the closed interval* $[a, b]$ can be proved by analogy. Make note of the fact that in this case we can take a as the lower limit of integration c.

Remark 2. While proving Theorem 10.6, we established the existence of a derivative of an integral with a variable upper limit and proved that derivative to be equal to the integrand function

$$\frac{d}{dx}\left(\int_c^x f(t)\,dt\right) = f(x). \tag{10.17}$$

Remark 3. It should be emphasized that if the function $f(x)$ is integrable on any closed interval contained in the interval (a, b), then the integral with a variable upper limit is a function of the upper limit continuous on the interval (a, b). To verify this fact, we shall prove that the increment $\Delta F = F(x + \Delta x) - F(x)$ of the function $F(x) = \int_c^x f(t)\,dt$ tends to zero as $\Delta x \to 0$. By virtue of formula (10.12) we have

$$\Delta F = F(x+\Delta x) - F(x) = \int_x^{x+\Delta x} f(t)\,dt = \mu \Delta x,$$

where the number μ is contained between the supremum and the infimum of the function $f(x)$ on the closed interval $[x, x + \Delta x]$. It follows from the last formula that ΔF also tends to zero as $\Delta x \to 0$.

Remark 4. An integral with a variable upper limit is often used to determine new functions. As we have noted in Chapter 6, the antiderivatives of some elementary functions cannot be expressed in terms of elementary functions and, consequently, are not elementary functions. Recall that functions $\int_0^x e^{-t^2}dt$, $\int_0^x \cos t^2\,dt$, for instance, are nonelementary functions.

10.7.2. The fundamental formula of integral calculus. We have proved that any two antiderivatives of the given function $f(x)$ differ by a constant (see Theorem 6.1). Therefore, in accordance with Theorem 10.6 and Remark 1 to that theorem, we can assert that any antiderivative $\Phi(x)$ of the function $f(x)$ continuous on the closed interval $[a, b]$ has the form

$$\Phi(x) = \int_a^x f(t)\,dt + C,$$

where C is some constant.

Putting first $x = a$ and then $x = b$ in the last formula and making use of property 1° of definite integrals, we shall find

$$\Phi(a) = C,\ \Phi(b) = \int_a^b f(x)\,dx + C^*.$$

These equalities yield the relation

$$\int_a^b f(x)\,dx = \Phi(b) - \Phi(a) \tag{10.18}$$

called the *fundamental formula of integral calculus***.

Thus, *to calculate the definite integral of a continuous function $f(x)$, it is necessary to compile the difference between the values of its arbitrary antiderivative for the upper and lower limits of integration.*

It should be pointed out that the fundamental formula of integral calculus offers strong possibilities of computing definite integrals, since the problem of computing the definite integral reduces to the problem of finding the antiderivative of a function. The methods of searching antiderivatives were given a thorough treatment in Chapters 6 and 7 of the present course.

Inasmuch as in many cases it is very difficult to find antiderivatives, it is natural to consider the question of the methods of approximation of definite integrals. Some such methods will be discussed in Chap. 12.

Formula (10.18) is sometimes given another form, namely, the difference $\Phi(b) - \Phi(a)$ is symbolized as $\Phi(x)\,|_a^b$. Then

$$\int_a^b f(x)\,dx = \Phi(x)\,|_a^b \tag{10.19}$$

Let us consider some examples.

(1) $\displaystyle\int_c^b \sin x\,dx = -\cos x\Big|_a^b = \cos a - \cos b,$

(2) $\displaystyle\int_1^2 \frac{dx}{x} = \ln x\Big|_1^2 = \ln 2 - \ln 1 = \ln 2,$

(3) $\displaystyle\int_0^1 e^{-x}\,dx = -e^{-x}\Big|_0^1 = 1 - \frac{1}{e},$

---

* We have used the letter x to designate the variable of integration in this formula since the upper limit has a fixed value b.

** This formula is also known as the formula of Newton-Leibniz.

$$(4)\quad \int_0^1 \frac{dx}{1+x^2} = \operatorname{arc\,tan} x \Big|_0^1 = \frac{\pi}{4},$$

$$(5)\quad \int_0^{\frac{1}{2}} \frac{dx}{\sqrt{1-x^2}} = \operatorname{arc\,sin} x \Big|_0^{\frac{1}{2}} = \frac{\pi}{6},$$

$$(6)\quad \int_0^3 \frac{dx}{\sqrt{1+x^2}} = \ln\left(x+\sqrt{1+x^2}\right)\Big|_0^3 = \ln\left(3+\sqrt{10}\right).$$

10.7.3. A change of variable under the definite integral sign. Suppose the following conditions are satisfied:

(1) *the function* $f(x)$ *is continuous on the closed interval* $[a, b]$.

(2) *the closed interval* $[a, b]$ *is the set of values of a certain function* $x = g(t)$ *defined on the closed interval* $\alpha \leqslant t \leqslant \beta$ *and possessing a continuous derivative on that interval*;

(3) $g(\alpha) = a$, $g(\beta) = b$.

Under these conditions the formula

$$\int_a^b f(x)\,dx = \int_\alpha^\beta f[g(t)]\,g'(t)\,dt \tag{10.20}$$

holds true. This formula shows that if the integral on the left-hand side has been calculated, then the integral on the right-hand side has been calculated too, and vice versa. The indicated formula is known as the *formula of a change of variable under the definite integral sign.*

Let us consider an antiderivative $\Phi(x)$ of the function $f(x)$. Formula (10.18) yields

$$\int_a^b f(x)\,dx = \Phi(b) - \Phi(a). \tag{10.21}$$

Since the functions $\Phi(x)$ and $x = g(t)$ are differentiable on the respective closed intervals, the composite function $\Phi(g(t))$ is differentiable on the closed interval $[\alpha, \beta]$. Therefore, applying the rule for differentiating a composite function, we obtain

$$\frac{d}{dt}\Phi(g(t)) = \Phi'(g(t))\,g'(t). \tag{10.22}$$

The derivative Φ' appearing in this formula is calculated from the argument x: $\Phi'(g(t)) = \Phi'(x)$, where $x = g(t)$. Since $\Phi'(x) = f(x)$, we obtain $\Phi'(g(t)) = f(g(t))$ at $x = g(t)$. Substituting this value

$\Phi'(g(t))$ into the right-hand side of equality (10.11), we obtain

$$\frac{d}{dt}\Phi(g(t)) = f(g(t))\,g'(t).$$

Consequently, the function $\Phi(g(t))$, defined on the closed interval $[\alpha, \beta]$, is, on that interval, an antiderivative of the function $f(g(t))\,g'(t)$, and, therefore, in accordance with formula (10.18),

$$\int_{\alpha}^{\beta} f(g(t))\,g'(t)\,dt = \Phi(g(\beta)) - \Phi(g(\alpha)).$$

Since $g(\beta) = b$ and $g(\alpha) = a$, we have

$$\int_{\alpha}^{\beta} f(g(t))\,g'(t)\,dt = \Phi(b) - \Phi(a).$$

Comparing the last formula with formula (10.21), we ascertain the validity of formula (10.20).

Examples. (1) Let us consider the integral $\int_1^2 \ln x \frac{dx}{x}$. We put $x = e^t$. Since $t = 0$ for $x = 1$, $t = \ln 2$ for $x = 2$, it follows that

$$\int_1^2 \ln x \frac{dx}{x} = \int_0^{\ln 2} t\,dt = \frac{t^2}{2}\Big|_0^{\ln 2} = \frac{1}{2}\ln^2 2.$$

(2) Let us now consider the integral $\int_{\pi^2/4}^{\pi^2} \sin\sqrt{x}\,\frac{dx}{\sqrt{x}}$. Suppose $x = t^2$. Then $x = \frac{\pi^2}{4}$ for $t = \frac{\pi}{2}$ and $x = \pi^2$ for $t = \pi$. Therefore

$$\int_{\frac{\pi^2}{4}}^{\pi^2} \sin\sqrt{x}\,\frac{dx}{\sqrt{x}} = 2\int_{\frac{\pi}{2}}^{\pi} \sin t\,dt = -2\cos t\Big|_{\frac{\pi}{2}}^{\pi} = 2.$$

10.7.4. The formula of integration by parts. Assume that the functions $u(x)$ and $v(x)$ possess continuous derivatives on the closed interval $[a, b]$. Then the following formula of integration by parts holds true for definite integrals:

$$\int_a^b u(x)\,v'(x)\,dx = [u(x)\,v(x)]\Big|_a^b - \int_a^b v(x)\,u'(x)\,dx. \qquad (10.23)$$

Since $v'(x)\,dx = dv$ and $u'(x)\,dx = du$, this formula can be written in the following form:

$$\int_a^b u\,dv = [uv]\Big|_a^b - \int_a^b v\,du \tag{10.24}$$

It is easy enough to ascertain the validity of these formulas. Indeed, the function $u(x)\,v(x)$ is an antiderivative of the function $u(x)\,v'(x) + v(x)\,u'(x)$. Therefore, by virtue of (10.19), we have

$$\int_a^b [u(x)\,v'(x) + v(x)\,u'(x)]\,dx = [u(x)\,v(x)]\Big|_a^b,$$

whence we can obtain formulas (10.23) and (10.24) by making use of property 3° of definite integrals (see 10.5).

Examples.

(1) $\displaystyle\int_1^2 \ln x\,dx = x\ln x\Big|_1^2 - \int_1^2 x\,\frac{dx}{x} = [x\ln x - x]\Big|_1^2 = 2\ln 2 - 1,$

(2) $\displaystyle\int_1^2 xe^x\,dx = xe^x\Big|_1^2 - \int_1^2 e^x\,dx = e^x(x-1)\Big|_1^2 = e^2,$

(3) $\displaystyle\int_0^1 \arctan x\,dx = x\arctan x\Big|_0^1 - \int_0^1 \frac{x\,dx}{1+x^2} =$

$$= \left[x\arctan x - \frac{1}{2}\ln(1+x^2)\right]\Big|_0^1 = \frac{\pi}{4} - \ln\sqrt{2}.$$

10.7.5. The remainder in Taylor's formula in integral form. Let us apply formula (10.23) for deriving *Taylor's formula for the function $f(x)$ with the remainder in integral form.* Assume that the function $f(x)$ possesses a continuous derivative of the $(n+1)$th order in some ε-neighbourhood of point a and suppose that x is an arbitrary given point belonging to that ε-neighbourhood. Let us make sure that the number

$$R_{n+1}(x) = \frac{1}{n!}\int_a^x f^{(n+1)}(t)\,(x-t)^n\,dt \tag{10.25}$$

is the remainder of Taylor's formula for the function $f(x)$ with the centre of expansion at point a. *Thus, formula* (10.25) *gives the representation of the remainder of Taylor's formula for the function $f(x)$ in integral form.*

To prove this, make note of the fact that

$$f(x) = f(a) + \int_a^x f'(t)\,dt.$$

We shall apply to the integral $\int_a^x f'(t)\,dt$ formula (10.23) of integration by parts putting $u(t) = f'(t)$ and $v(t) = -(x-t)$ ($v'\,dt = dt$ since x is a fixed value). We have

$$\int_a^x f'(t)\,dt = -f'(t)(x-t)\Big|_a^x + \int_a^x f''(t)(x-t)\,dt =$$

$$= f'(a)(x-a) + \int_a^x f''(t)(x-t)\,dt.$$

Substituting the expression obtained for $\int_a^x f'(t)\,dt$ into the above-written formula for $f(x)$, we arrive at

$$f(x) = f(a) + f'(a)(x-a) + \int_a^x f''(t)(x-t)\,dt.$$

We can also apply the formula of integration by parts to the integral $\int_a^x f''(t)(x-t)\,dt$, putting $u(t) = f''(t)$ and $v(t) = -\frac{1}{2}(x-t)^2$ ($v'\,dt = (x-t)\,dt$ since x is a fixed value. After a simple transformation we obtain

$$\int_a^x f''(t)(x-t)\,dt = \frac{f''(a)}{2!}(x-a)^2 + \frac{1}{2!}\int_a^x f^{(3)}(t)(x-t)^2\,dt$$

and, therefore,

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{1}{2!}\int_a^x f^{(3)}(t)(x-t)\,dt.$$

We shall proceed with integrating by parts until we arrive at the formula

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots +$$

$$+ \frac{f^{(n)}(a)}{n!}(x-a)^n + \frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n\,dt.$$

This formula shows that $R_{n+1}(x)$ is really the remainder in Taylor's formula for the function $f(x)$ with centre of expansion at point a (see 8.13). Using the integral form (10.25) of the remainder of Taylor's formula, we can easily obtain the remainder of Taylor's formula in Lagrange's form. Indeed, we receive from the generalized form (10.15) of the mean-value formula

$$R_{n+1}(x)=\frac{1}{n!}\int_a^x f^{(n+1)}(t)\,(x-t)^n\,dt=\frac{f^{(n+1)}(\xi)}{n!}\int_a^x (x-t)^n\,dt=$$

$$=-\frac{f^{(n+1)}(\xi)}{n!}\,\frac{(x-t)^{n+1}}{(n+1)}\Big|_a^x=\frac{f^{(n+1)}(\xi)}{(n+1)!}\,(x-a)^{n+1}.$$

The expression we have obtained is precisely the remainder in Lagrange's form* (see formula (8.46) in 8.14).

SUPPLEMENT 1 TO CHAPTER 10

SOME IMPORTANT INEQUALITIES FOR SUMS AND INTEGRALS

1.1. Derivation of a preliminary inequality. Assume that A and B are arbitrary nonnegative numbers and p and p' are two arbitrary numbers, both exceeding unity and connected by the relation $\frac{1}{p}+\frac{1}{p'}=1$ (they are known as conjugate numbers). Then

$$AB\leqslant\frac{A^p}{p}+\frac{B^{p'}}{p'}. \tag{10.26}$$

Let us find the maximum value of the function $f(x)=x^{1/p}-\frac{x}{p}$ on the half-line $x\geqslant 0$. Since $f'(x)=\frac{1}{p}\left(x^{\frac{1}{p}-1}-1\right)=\frac{1}{p}\left(x^{-\frac{1}{p'}}-1\right)$, we have $f'(x)>0$ for $0<x<1$ and $f'(x)<0$ for $x>1$. Therefore, the function possesses a maximum at point $x=1$ and its maximum value $f(1)=1-\frac{1}{p}=\frac{1}{p'}$. Thus for all $x\geqslant 0$

$$x^{\frac{1}{p}}-\frac{x}{p}\leqslant\frac{1}{p'}.$$

* Pay attention to the fact that in the indicated derivation of the remainder in Lagrange's form, the restrictions imposed on the derivative of the $(n+1)$-th order are somewhat greater than those indicated in 8.14. However, if we make use of Darboux's theorem (on the passage of the derivative through all the intermediate values) we have proved at the end of Chapter 9, we shall obtain the remainder in Lagrange's form only under the condition of the existence and integrability of $f^{(n+1)}(x)$.

Having fixed $x = \frac{A^p}{B^{p'}}$ * in the last inequality and multiplied both sides of this inequality by $B^{p'}$, we obtain inequality (10.26).

1.2. **Hölder's** ** **inequality for sums.** Assume $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ be any nonnegative numbers and p and p' have the same meaning as before. Then the following inequality is valid:

$$\sum_{i=1}^{n} a_i b_i \leqslant \left[\sum_{i=1}^{n} a_i^p\right]^{\frac{1}{p}} \left[\sum_{i=1}^{n} b_i^{p'}\right]^{\frac{1}{p'}} \tag{10.27}$$

which is known as *Hölder's inequality for sums.*

We shall first prove that if $A_1, A_2, \ldots, A_n$; $B_1, B_2, \ldots, B_n$ are arbitrary nonnegative numbers satisfying the inequalities

$$\sum_{i=1}^{n} A_i^p \leqslant 1 \quad \text{and} \quad \sum_{i=1}^{m} B_i^{p'} \leqslant 1, \tag{10.28}$$

these numbers also comply with the inequality

$$\sum_{i=1}^{n} A_i B_i \leqslant 1. \tag{10.29}$$

Indeed, writing inequalities (10.26) for all pairs of numbers A_i and B_i and summing up these inequalities over all i from 1 to n, we obtain

$$\sum_{i=1}^{n} A_i B_i \leqslant \frac{1}{p} \sum_{i=1}^{n} A_i^p + \frac{1}{p'} \sum_{i=1}^{n} B_i^{p'} \leqslant \frac{1}{p} + \frac{1}{p'} = 1.$$

We have thus proved inequality (10.29).

Let us now put

$$A_i = \frac{a_i}{\left[\sum_{i=1}^{n} a_i^p\right]^{\frac{1}{p}}}, \quad B_i = \frac{b_i}{\left[\sum_{i=1}^{n} b_i^{p'}\right]^{\frac{1}{p'}}}. \text{ ***}$$

It is easy to see that the numbers A_i and B_i satisfy inequalities (10.28) and, therefore, inequality (10.29) is valid for these numbers. In the given case, this inequality can be written as

$$\frac{\sum_{i=1}^{n} a_i b_i}{\left[\sum_{i=1}^{n} a_i^p\right]^{\frac{1}{p}} \left[\sum_{i=1}^{n} b_i^{p'}\right]^{\frac{1}{p'}}} \leqslant 1.$$

The last inequality yields Hölder's inequality (10.27).

---

* Here we assume that $B > 0$ since for $B = 0$ the validity of inequality (10.26) is obvious.

** Otto Hölder (1859-1937), a German mathematician.

*** We assume at least one of the numbers a_i and at least one of the numbers b_i to be nonzero, since otherwise formula (10.27) does not require to be proved.

Remark. In a particular case $p = p' = 2$, Hölder's inequality passes into the following inequality:

$$\sum_{i=1}^{n} a_i b_i \leqslant \sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}. \tag{10.30}$$

Inequality (10.30) is called *Buniakovsky's* inequality for sums.*

1.3. Minkowski's inequality for sums.** Suppose that $a_1, a_2, \ldots, a_n$; $b_1, b_2, \ldots, b_n$ are arbitrary nonnegative numbers and the number $p > 1$. Then the following inequality is valid:

$$\left[\sum_{i=1}^{n} (a_i + b_i)^p\right]^{\frac{1}{p}} \leqslant \left[\sum_{i=1}^{n} a_i^p\right]^{\frac{1}{p}} + \left[\sum_{i=1}^{n} b_i^p\right]^{\frac{1}{p}} \tag{10.31}$$

which is called *Minkowski's inequality for sums.* Let us first transform the sum on the left-hand side of (10.31). We can write

$$\sum_{i=1}^{n} (a_i + b_i)^p = \sum_{i=1}^{n} a_i (a_i + b_i)^{p-1} + \sum_{i=1}^{n} b_i (a_i + b_i)^{p-1}.$$

We shall apply Hölder's inequality to every sum on the right-hand side and obtain, since $(p-1)\, p' = p$ and $\frac{1}{p'} = \frac{p-1}{p}$,

$$\sum_{i=1}^{n} (a_i + b_i)^p \leqslant \left[\sum_{i=1}^{n} a_i^p\right]^{\frac{1}{p}} \left[\sum_{i=1}^{n} (a_i + b_i)^{(p-1)p'}\right]^{\frac{1}{p'}} +$$

$$+ \left[\sum_{i=1}^{n} b_i^p\right]^{\frac{1}{p}} \left[\sum_{i=1}^{n} (a_i + b_i)^{(p-1)p'}\right]^{\frac{1}{p'}} =$$

$$= \left\{\left[\sum_{i=1}^{n} a_i^p\right]^{\frac{1}{p}} + \left[\sum_{i=1}^{n} b_i^p\right]^{\frac{1}{p}}\right\} \left[\sum_{i=1}^{n} (a_i + b_i)^p\right]^{\frac{p-1}{p}}.$$

Having divided both sides of the last inequality by $\left[\sum_{i=1}^{n} (a_i + b_i)^p\right]^{\frac{p-1}{p}}$, we obtain Minkowski's inequality (10.31).

1.4. Integrability of an arbitrary positive degree of the modulus of the function being integrated. Let us prove the following theorem.

Theorem 10.7. *If the function $f(x)$ is integrable on the closed interval $[a, b]$ then the function $|f(x)|^r$, where r is any positive real number, is also integrable on the interval $[a, b]$.*

Proof. It is sufficient to prove the theorem for the case $r < 1$, since in the case $r > 1$ the function $|f(x)|^p$ can be represented as the product $|f(x)|^{[r]} \times |f(x)|^{r-[r]}$, where $[r]$ is the integral part of r and $r - [r] < 1$. By virtue of Remark 2 made in 6.1, the function $|f(x)|$ is integrable on the closed interval $[a, b]$ and, therefore, by virtue of property 3° of Sec. 5, the function $|f(x)|^{[r]}$ is integrable on that interval. But then, in accordance with the same

* V. Ya. Buniakovsky (1804-1889), a Russian mathematician.

** Hermann Minkowski (1864-1909), a German mathematician and physicist.

property and the integrability of the function $|f(x)|^{r-[r]}$, the function $|f(x)|^r$ is also integrable on the closed interval $[a, b]$. So we shall prove the theorem for the case $r < 1$. We shall put $r = \frac{1}{p}$ and note that $p > 1$. Since the function $|f(x)|$ is integrable on the interval $[a, b]$ there is a partition T of that interval for any $\varepsilon > 0$, such that

$$\sum_{i=1}^{n} (M_i - m_i)\,\Delta x_i < \varepsilon^p (b-a)^{(1-p)}. \tag{10.32}$$

Here M_i and m_i denote the supremum and the infimum of the function $|f(x)|$ on the subinterval $[x_{i-1}, x_i]$. It is sufficient to prove that the sum

$$S - s = \sum_{i=1}^{n} \left(M_i^{\frac{1}{p}} - m_i^{\frac{1}{p}} \right) \Delta x_i \tag{10.33}$$

is smaller than ε.

Let us use Hölder's inequality (10.27) to evaluate this sum, putting in it $a_i = (M_i^{1/p} - m_i^{1/p})\,(\Delta x_i)^{1/p}$, $b_i = (\Delta x_i)^{1/p'}$. We obtain

$$S - s \leqslant \left[\sum_{i=1}^{n} \left(M_i^{\frac{1}{p}} - m_i^{\frac{1}{p}} \right)^p \Delta x_i \right]^{\frac{1}{p}} \left[\sum_{i=1}^{n} \Delta x_i \right]^{\frac{1}{p'}} \tag{10.34}$$

Let us now prove that

$$\left(M_i^{\frac{1}{p}} - m_i^{\frac{1}{p}} \right)^p \leqslant (M_i - m_i). \tag{10.35}$$

Being divided by M_i*, the last inequality reduces to the following inequality:

$$\left[1 - \left(\frac{m_i}{M_i} \right)^{\frac{1}{p}} \right]^p \leqslant 1 - \frac{m_i}{M_i}.$$

We can easily ascertain the validity of the last inequality if we take into account that $0 \leqslant \frac{m_i}{M_i} \leqslant 1$ and $p > 1$. Making use of inequality (10.35) and taking into consideration that

$$\sum_{i=1}^{n} \Delta x_i = b - a,$$

we shall obtain from (10.34) the following inequality:

$$S - s \leqslant \left[\sum_{i=1}^{n} (M_i - m_i)\,\Delta x_i \right]^{\frac{1}{p}} (b-a)^{\frac{1}{p'}}.$$

Hence, applying inequality (10.32) and taking into account that $\frac{1}{p} + \frac{1}{p'} = 1$, we shall have

$$S - s < \varepsilon.$$

The theorem is proved.

1.5. Hölder's inequality for integrals. Suppose $f(x)$ and $g(x)$ are any two functions integrable on the closed interval $[a, b]$, and p and p' are any two

* We can assume that $M_i > 0$, since if $M_i = 0$, we have $m_i = 0$, too, and inequality (10.35) holds true.

numbers, which both exceed unity and are connected by the relation $\frac{1}{p}+\frac{1}{p'}=1$. Then the following inequality holds true:

$$\left|\int_a^b f(x)\,g(x)\,dx\right| \leqslant \left[\int_a^b |f(x)|^p\,dx\right]^{\frac{1}{p}} \left[\int_a^b |g(x)|^{p'}\,dx\right]^{\frac{1}{p'}}, \tag{10.36}$$

which is called *Hölder's inequality for integrals*. Make note of the fact that the existence of integrals on the right-hand side of inequality (10.36) is ensured by Theorem 10.7, and of the integral on the left-hand side by property 3° in Sec. 5.

Let us first prove that if $A(x)$ and $B(x)$ are two nonnegative functions integrable on the closed interval $[a, b]$ and satisfying the inequalities

$$\int_a^b A^p(x)\,dx \leqslant 1, \quad \int_a^b B^{p'}(x)\,dx \leqslant 1, \tag{10.37}$$

then

$$\int_a^b A(x)\,B(x)\,dx \leqslant 1. \tag{10.38}$$

Indeed, at any point x belonging to the closed interval $[a, b]$ inequality (10.26) holds true:

$$A(x)\,B(x) \leqslant \frac{A^p(x)}{p}+\frac{B^{p'}(x)}{p'}.$$

Hence, by virtue of evaluation 3° from 10.6 and formulas (10.37),

$$\int_a^b A(x)\,B(x)\,dx \leqslant \frac{1}{p}\int_a^b A^p(x)\,dx+\frac{1}{p'}\int_a^b B^{p'}(x)\,dx \leqslant \frac{1}{p}+\frac{1}{p'}=1.$$

We have thus proved inequality (10.38).

Assuming

$$A(x)=\frac{|f(x)|}{\left[\int_a^b |f(x)|^p\,dx\right]^{\frac{1}{p}}}, \quad B(x)=\frac{|g(x)|}{\left[\int_a^b |g(x)|^{p'}\,dx\right]^{\frac{1}{p'}}},$$

we arrive at the following inequality:

$$\int_a^b |f(x)|\,|g(x)|\,dx \leqslant \left[\int_a^b |f(x)|^p\,dx\right]^{\frac{1}{p}} \left[\int_a^b |g(x)|^{p'}\,dx\right]^{\frac{1}{p'}}.$$

Since by virtue of Remark 2 in 6.1,

$$\left|\int_a^b f(x)\,g(x)\,dx\right| \leqslant \int_a^b |f(x)|\,|g(x)|\,dx,$$

the proof of Hölder's inequality (10.36) for integrals is complete.

Remark. In a particular case $p = p' = 2$, Hölder's inequality for integrals passes into the following inequality:

$$\left| \int_a^b f(x)\, g(x)\, dx \right| \leqslant \sqrt{\int_a^b |f(x)|^2\, dx} \sqrt{\int_a^b |g(x)|^2\, dx}, \tag{10.39}$$

which is known as the *Cauchy-Buniakovsky inequality for integrals.*

1.6. Minkowski's inequality for integrals. For any functions $f(x)$ and $g(x)$ which are nonnegative and integrable on the closed interval $[a, b]$ and for any number $p > 1$ the following inequality holds true:

$$\left\{ \int_a^b [f(x)+g(x)]^p\, dx \right\}^{\frac{1}{p}} \leqslant \left[\int_a^b f^p(x)\, dx \right]^{\frac{1}{p}} + \left[\int_a^b g^p(x)\, dx \right]^{\frac{1}{p}}, \tag{10.40}$$

called *Minkowski's inequality for integrals.* To obtain this inequality we must proceed from the formula

$$\int_a^b [f(x)+g(x)]^p\, dx = \int_a^b f(x)\, [f(x)+g(x)]^{p-1}\, dx + \int_a^b g(x)\, [f(x)+g(x)]^{p-1}\, dx$$

and apply Hölder's inequality to the integrals on the right-hand side of this formula. We leave the details of the argument to the reader.

By the method of induction, we can obtain from (10.40) the following inequality for the n functions $f_1(x)$, $f_2(x)$, ..., $f_n(x)$, nonnegative and integrable on the closed interval $[a, b]$:

$$\left\{ \int_a^b [f_1(x)+f_2(x)+\ldots+f_n(x)]^p\, dx \right\}^{\frac{1}{p}} \leqslant$$

$$\leqslant \left[\int_a^b f_1^p(x)\, dx \right]^{\frac{1}{p}} + \left[\int_a^b f_2^p(x)\, dx \right]^{\frac{1}{p}} + \ldots + \left[\int_a^b f_n^p(x)\, dx \right]^{\frac{1}{p}}.$$

SUPPLEMENT 2 TO CHAPTER 10

PROOF OF THE ASSERTION FROM 10.6.4

For the sake of convenience, let us formulate once again the assertion made in 10.6.4.

If the function $g(x)$ is monotonic and the function $f(x)$ is integrable on the closed interval $[a, b]$, then there is a number ξ on that interval such that

$$\int_a^b f(x)\, g(x)\, dx = g(a) \int_a^\xi f(x)\, dx + g(b) \int_\xi^b f(x)\, dx. \tag{10.16*}$$

We shall first prove the following auxiliary proposition.

* For the sake of convenience, we retain the numeration of the formula given in 10.6.4.

Abel's* Lemma. *Suppose* $v_1 \geqslant v_2 \geqslant \ldots \geqslant v_n \geqslant 0$ *and* $u_1, u_2, \ldots, u_n$ *are arbitrary numbers. If the sums* $S_i = u_1 + u_2 + \ldots + u_i$ *are contained between* A *and* B *for any* i, *then the sum* $v_1u_1 + v_2u_2 + \ldots + v_nu_n$ *is contained between the numbers* Av_1 *and* Bv_1.

Proof. We have $u_1 = S_1$ and $u_i = S_i - S_{i-1}$. Therefore,

$$v_1u_1 + v_2u_2 + \ldots + v_nu_n = v_1S_1 + v_2(S_2 - S_1) + \ldots + v_n(S_n - S_{n-1}) =$$
$$= S_1(v_1 - v_2) + S_2(v_2 - v_3) + \ldots + S_{n-1}(v_{n-1} - v_n) + S_nv_n.$$

Since $v_i \geqslant 0$ and $v_i - v_{i+1} \geqslant 0$, we can replace in the last relation each S_i first by A and then by B and obtain inequalities

$$A[(v_1 - v_2) + (v_2 - v_3) + \ldots + (v_{n-1} - v_n) + v_n] \leqslant$$
$$\leqslant v_1u_1 + v_2u_2 + \ldots + v_nu_n \leqslant B[(v_1 - v_2) + (v_2 - v_3) + \ldots +$$
$$+ (v_{n-1} - v_n) + v_n].$$

Hence, paying attention to the fact that the expressions in square brackets are equal to v_1, we obtain

$$Av_1 \leqslant v_1u_1 + v_2u_2 + \ldots + v_nu_n \leqslant Bv_1.$$

We have proved the lemma.

Remark. While proving Abel's lemma we made use of a transformation of the sum $\sum_{k=1}^{n} v_ku_k$ which is usually called *Abel's transformation.* A more detailed discussion of Abel's transformation and its important applications can be found in 13.5.2.

Proof of the assertion made in 10.6.4. Let us assume that the function $g(x)$ does not increase on $[a, b]$ and is nonnegative on that interval. Then we have, by virtue of the integrability of $f(x)\,g(x)$**,

$$\int_a^b f(x)g(x)\,dx = \lim_{\Delta \to 0} \sum_{i=1}^{n} f(x_{i-1})\,g(x_{i-1})\,\Delta x_i, \quad \text{where } \Delta = \max \Delta x_i.$$

Suppose now that M_i and m_i are the supremum and the infimum of $f(x)$ on the closed interval $[x_{i-1}, x_i]$. Then, since $g(x)$ is nonnegative, the following inequalities hold true:

$$\sum_{i=1}^{n} m_ig(x_{i-1})\Delta x_i \leqslant \sum_{i=1}^{n} f(x_{i-1})\,g(x_{i-1})\,\Delta x_i \leqslant \sum_{i=1}^{n} M_ig(x_{i-1})\Delta x_i. \qquad (10.41)$$

Since $g(x)$ does not increase on $[a, b]$, the difference

$$\sum_{i=1}^{n} M_ig(x_{i-1})\Delta x_i - \sum_{i=1}^{n} m_ig(x_{i-1})\Delta x_i = \sum_{i=1}^{n} (M_i - m_i)\,g(x_{i-1})\,\Delta x_i$$

does not exceed the number $g(a)\sum_{i=1}^{n}(M_i - m_i)\,\Delta x_i$. Since the function $f(x)$ is integrable, the sum $\sum_{i=1}^{n}(M_i - m_i)\,\Delta x_i = \sum_{i=1}^{n} \omega_i\,\Delta x_i$ tends to zero as $\Delta \to 0$. It follows from this and from inequalities (10.41) that for *any numbers* μ_i

* Niels Henrik Abel (1802-1 29), a Norwegian mathematician.

** See Property 3° in 10.5.

satisfying the inequalities $m_i \leqslant \mu_i \leqslant M_i$ each of the sums

$$\sum_{i=1}^{n} m_i g(x_{i-1})\,\Delta x_i, \quad \sum_{i=1}^{n} \mu_i g(x_{i-1})\,\Delta x_i, \quad \sum_{i=1}^{n} M_i g(x_{i-1})\,\Delta x_i$$

has the integral $\int_a^b f(x)\,g(x)\,dx$ as its limit for $\Delta \to 0$. In accordance with formula (10.12), the numbers μ_i, $m_i \leqslant \mu_i \leqslant M_i$ can be chosen such that $\int_{x_{i-1}}^{x_i} f(x)\,dx = \mu_i\,\Delta x_i$. The function $F(x) = \int_a^x f(t)\,dt$ being continuous on the closed interval $[a, b]$ (see Remark 3 in 7.1), the numbers $S_i = \sum_{k=1}^{i} \mu_k\,\Delta x_k = \int_a^{x_i} f(t)\,dt$ lie between the infimum m and the supremum M of the function $F(x)$ on the interval $[a, b]$. We shall put $v_1 = g(a)$, $v_2 = g(x_1)$, ..., $v_n = g(x_{n-1})$, $u_1 = \mu_1\,\Delta x_1$, ..., $u_n = \mu_n\,\Delta x_n$. Since $v_1 \geqslant v_2 \geqslant \ldots \geqslant v_n \geqslant 0$ and the sums $S_i = \sum_{k=1}^{i} u_k$ lie between m and M, it follows, by virtue of Abel's lemma, that the sum $\sum_{i=1}^{n} g(x_{i-1})\,\mu_i\,\Delta x_i$ lies between $mg(a)$ and $Mg(a)$. But in that case, the limit of this sum, for $\Delta \to 0$, lies between $mg(a)$ and $Mg(a)$, i.e. the inequalities

$$g(a)\,m \leqslant \int_a^b f(x)\,g(x)\,dx \leqslant g(a)\,M$$

hold true. The continuous function $F(x) = \int_a^x f(t)\,dt$ assumes any value of A contained between its infimum m and supremum M, that is, there is a point ξ such that

$$F(\xi) = \int_a^{\xi} f(t)\,dt = A = \frac{\int_a^b f(x)\,g(x)\,dx}{g(a)}.$$

Therefore,

$$\int_a^b f(x)\,g(x)\,dx = g(a) \int_a^{\xi} f(x)\,dx. \tag{10.42}$$

If the nonincreasing function $g(x)$ possesses negative values too, then the function $h(x) = g(x) - g(b)$ is nonincreasing and possesses nonnegative values. Therefore, by virtue of (10.42), we have

$$\int_a^b f(x)\,[g(x) - g(b)]\,dx = [g(a) - g(b)] \int_a^{\xi} f(x)\,dx,$$

whence, by means of simple transformations, we obtain formula (10.16).

CHAPTER 11

APPLICATIONS OF THE DEFINITE INTEGRAL IN GEOMETRY AND PHYSICS

11.1. THE LENGTH OF THE ARC OF A CURVE

11.1.1. The concept of a plane curve. It is most natural to consider a curve as the trace of a moving point. Here we shall impart a precise geometrical meaning to the concept of a curve and introduce the notion of the so-called *simple curve.*

Suppose the functions $\varphi(t)$ and $\psi(t)$ are continuous on the closed interval $[\alpha, \beta]$ (in what follows the argument of these functions will be called a parameter). If we assume the parameter t to represent time, the indicated functions will determine the law of motion of point M with coordinates

$$x = \varphi(t), \quad y = \psi(t), \quad \alpha \leqslant t \leqslant \beta \tag{11.1}$$

along a plane (Fig. 11.1)*. The set $\{M\}$ of points M corresponding to various values of parameter t belonging to the closed interval $[\alpha, \beta]$ is naturally treated as the *trace* of point M moving in accordance with the law (11.1). Pay attention to the fact that the set $\{M\}$ which is the trace of a moving point may not be in keeping with our customary conception of a curve. There are, for instance, continuous functions $\varphi(t)$ and $\psi(t)$, specified on the closed interval $[0, 1]$, such that the trace of point M moving according to the law $x = \varphi(t)$, $y = \psi(t)$ will fill up a whole square**. So it is natural to separate the sets $\{M\}$ which correspond to our visual conception of a curve. Thus we come to the concept of a *simple curve.*

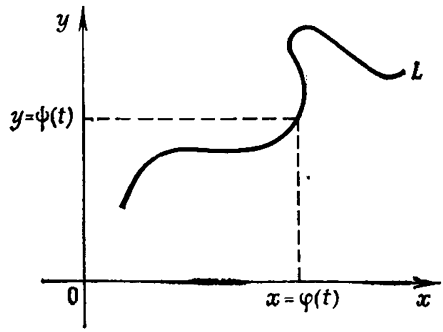


Fig. 11.1

The set $\{M\}$ of all the points M whose coordinates x and y are specified by equations (11.1) *will be called a simple plane curve L, if various values of parameter t belonging to the closed interval $[\alpha, \beta]$ are associated with various points of that set.*

---

* Here and in what follows a plane means a totality of various ordered pairs (x, y) of numbers x and y (every such pair will be called a point of the plane). The numbers x and y are the coordinates of point (x, y). For brevity, we shall also denote point (x, y) by the letter M. The notation $M(x, y)$ signifies that point M has the coordinates x and y.

** See L. Bieberbach, *Differential und Integralrechnung*, Bd. 1, s. 107.

We shall use the following phraseology: "*equations* (11.1) *specify the simple plane curve L*" and "*a simple plane curve L is represented parametrically* (*parametrized*) *by means of equations* (11.1)".

Every point of the set $\{M\}$ appearing in the definition of a simple plane curve will be called a point of that curve, with the points corresponding to the boundary values α and β of parameter t being called the boundary points of a simple curve.

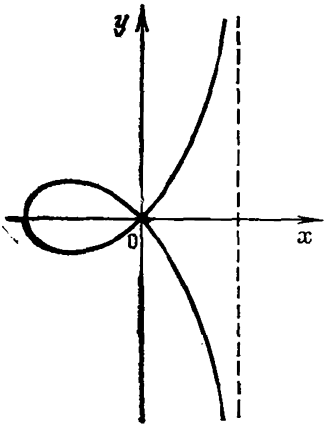


Fig. 11.2

The graph of a continuous function $y = f(x)$ belonging to the closed interval $[\alpha, \beta]$ can serve as an example of a simple curve. Indeed, this graph can be treated as the trace of point M moving in accordance with the law $x = t$, $y = f(t)$, $\alpha \leqslant t \leqslant \beta$, with various points of the graph evidently corresponding to various values of parameter t.

Remark 1. Simple curves do not cover all the point sets which can be called curves. For our aims, however, the notion of a simple curve is quite sufficient.

Remark 2. Different methods can be used to parametrize one and the same simple curve L. We shall consider various parametric representations of a simple curve L which can be obtained from a given parametrization by representing parameter t as continuous strictly monotonic functions of another parameter s.

Remark 3. An important concept is that of a *simple closed curve*. Such a curve can be formed in the following way. Suppose L_1 and L_2 are two simple curves, with (1) the boundary points of curve L_1 coinciding with the boundary points of curve L_2; (2) any points not lying on the boundaries of curves L_1 and L_2 being distinct. A curve L obtained by combining the curves L_1 and L_2 is called a simple closed curve.

11.1.2. Parametric representation of a curve. In mathematical analysis and its applications it is convenient to consider curves represented parametrically. A visual source of such a method of representation of a curve is an *understanding of a curve as a locus of successive positions of a moving point*. For instance, the locus of successive positions of point M with coordinates x and y, moving in accordance with the law

$$x = a\frac{t^2-1}{t^2+1}, \quad y = at\frac{t^2-1}{t^2+1}, \quad -\infty < t < \infty \tag{11.2}$$

is a curve called a *strophoid* (Fig. 11.2). Note that point M moving along a strophoid gets to one and the same position $x = 0$, $y = 0$ twice, at $t = -1$ and $t = 1$. Since we are considering successive positions of a moving point, it is natural to assume the points of a

strophoid corresponding to various values of parameter t to be distinct.

A strophoid is not a simple curve. It is, however, easy to verify that the domain of parameter t can be divided into parts so that the corresponding parts of the strophoid will be simple curves. Let us, for example, divide the number line $-\infty < t < \infty$ into closed intervals $[n-1, n]$, where n is any integer. It is evident, that if we consider parameter t on this interval, the corresponding part of the strophoid will be a simple curve.

We shall make use of the idea of division into parts to define mathematically the notion of a curve represented parametrically.

We shall assume the set $\{t\}$ to be either a closed interval, or a half-open interval, or an open interval, a number line, or, again, an open or closed half-line.

It is necessary at this point to introduce the notion of a partition of the set $\{t\}$. We shall say that a finite or infinite system of closed intervals $\{[t_{i-1}, t_i]\}$ divides the set $\{t\}$ if (1) all the intervals combined constitute the entire set $\{t\}$ and (2) only the end points of any two intervals of the system can be their common points.

Let us consider several examples of partitioning some of the indicated sets $\{t\}$.

1. The system of closed intervals $\left[0, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{2}{3}\right], \left[\frac{2}{3}, 1\right]$ evidently divides the closed interval $[0, 1]$.

2. The system of closed intervals $\left[0, \frac{1}{2}\right], \left[\frac{1}{2}, \frac{3}{4}\right], \left[\frac{3}{4}, \frac{7}{8}\right], \ldots, \left[\frac{2^n-1}{2^n}, \frac{2^{n+1}-1}{2^{n+1}}\right], \ldots$ divides the half-open interval $[0, 1)$.

3. The system of closed intervals $\{[n-1, n]\}$, where n is any integer, evidently divides the entire number line.

Let us now define the notion of a curve represented parametrically.

Assume the functions $\varphi(t)$ and $\psi(t)$ to be continuous on the set $\{t\}$*. We shall say that *the equations*

$$x = \varphi(t), \quad y = \psi(t) \tag{11.3}$$

specify the curve L parametrically if there exists a system of closed intervals $\{[t_{i-1}, t_i]\}$, dividing the set $\{t\}$, such that for the values of t, belonging to each given interval of the system, equations (11.3) *specify a simple curve.*

In this case, *the points of curve L are considered in increasing order of parameter t,* to wit, if point M_1 corresponds to the value of parameter t_1 and point M_2 to the value t_2, then M_1 is regarded as preceding M_2 if $t_1 < t_2$. It should be pointed out that *the points corresponding to various values of the parameter are always considered to be different.*

* The set $\{t\}$ is one of the sets indicated above.

In other words, a curve represented parametrically can be regarded as a combination of simple curves, with these simple curves being run through successively by a point M whose coordinates are determined by relation (11.3) when parameter t monotonically runs through the set $\{t\}$.

Remark 1. A simple curve can be considered as a curve represented parametrically. In that case, the system of closed intervals dividing the interval $[\alpha, \beta]$ reduces to that one interval.

Let us consider, as an example, a curve L specified parametrically by the equations

$$x = \cos t, \quad y = \sin t, \tag{11.4}$$

where t varies on the closed interval $[0, 4\pi]$. The system of closed intervals $[0, \pi]$, $[\pi, 2\pi]$, $[2\pi, 3\pi]$, $[3\pi, 4\pi]$ evidently divides the closed interval $[0, 4\pi]$, and for each value of t from every indicated interval of the system equations (11.4) specify a simple curve (semicircle). It can be seen that in the example being considered the curve L is a circle run through twice.

Remark 2. The example we have discussed as well as the example of a strophoid show that a curve represented parametrically may have points of self-intersection and even whole areas of self-overlapping.

Remark 3. In the case of a curve specified parametrically by equations (11.3), we shall also speak of parametrization of the indicated curve by means of these equations. One and the same curve L can be parametrized by different methods. We shall consider various parametrizations of curve L obtained from any given parametrization by representing parameter t as continuous strictly increasing functions of another parameter s. It should be noted that only with these transformations of a parameter is the order of points on the curve L retained.

11.1.3. The notion of a space curve. The notion of a space curve is introduced by a complete analogy with the notion of a plane curve. First the concept of a simple space curve is introduced as a set $\{M\}$ of space points whose coordinates x, y and z are specified by the equations

$$x = \varphi(t), \quad y = \psi(t), \quad z = \chi(t), \quad \alpha \leqslant t \leqslant \beta \tag{11.5}$$

under the conditions of continuity of the functions $\varphi(t)$, $\psi(t)$, $\chi(t)$ and the condition of the noncoincidence of the points of the set $\{M\}$ corresponding to various values of parameter t.

As in the case of a plane curve, the notion of a simple space curve and that of a partition of the set $\{t\}$ of variation of the parameter lead to the concept of a space curve specified parametrically by equations (11.5) under the condition of a monotonic variation of parameter t on the set $\{t\}$.

Note that the terminology introduced in the previous subsections is naturally carried over to the case of space curves.

11.1.4. The concept of an arc length of a curve. We shall introduce here the concept of the arc length of a curve represented parametrically.

Suppose a curve L is specified parametrically by the equations

$$x = \varphi(t), \quad y = \psi(t), \tag{11.3}$$

where the parameter t varies on the closed interval $[\alpha, \beta]$.

Assume T to be an arbitrary partition of the interval $[\alpha, \beta]$ by the points $\alpha = t_0 < t_1 < t_2 < \ldots < t_n = \beta$. Let us denote by

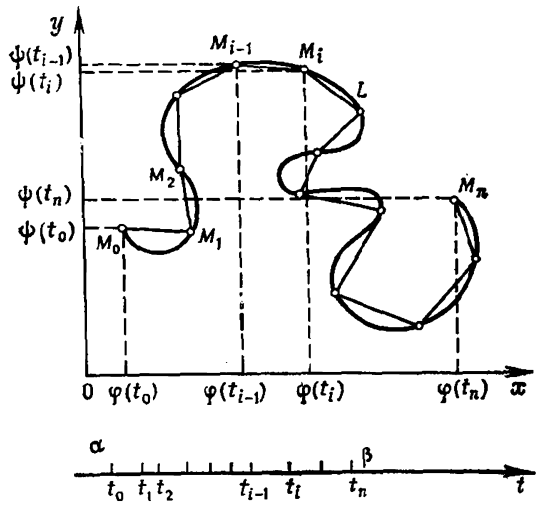


Fig. 11.3

$M_0, M_1, M_2, \ldots, M_n$ the respective points of the curve L (Fig. 11.3). The broken line $M_0M_1M_2 \ldots M_n$ appearing as a result will be called an open polygon* (a broken polygonal line) inscribed into the curve L and corresponding to the given partition T of the interval $[\alpha, \beta]$. Since the length $\overline{l}_i$ of the element $M_{i-1}M_i$ of that polygonal line is equal to

$$\sqrt{[\varphi(t_i) - \varphi(t_{i-1})]^2 + [\psi(t_i) - \psi(t_{i-1})]^2}\ **,$$

* We shall use the term a *straight line* for a line specified parametrically by the equations $x = at + b$, $y = ct + d$. The constants a, b, c and d can obviously be chosen such that the line would pass through two given points $M_1(x_1, y_1)$ and $M_2(x_2, y_2)$. It is natural to call the part of the line between points M_1 and M_2 a *segment* and the totality of a finite number of segments adjoining one another an *open polygon*.

** We have used the formula specifying the distance between the two points M_{i-1}, M_i, whose coordinates are equal, respectively, to

$$x_{i-1} = \varphi(t_{i-1}), \quad y_{i-1} = \psi(t_{i-1}) \text{ and } x_i = \varphi(t_i), \quad y_i = \psi(t_i).$$

the length $\bar{l}(t_i)$ of the whole broken line is

$$\bar{l}(t_i) = \sum_{i=1}^{n} \bar{l}_i = \sum_{i=1}^{n} \sqrt{[\varphi(t_i) - \varphi(t_{i-1})]^2 + [\psi(t_i) - \psi(t_{i-1})]^2}. \quad (11.6)$$

Definition. *If the set* $\{\bar{l}(t_i)\}$ *of the lengths of open polygons, inscribed into the curve* L *and corresponding to various partitions* T *of the closed interval* $[\alpha, \beta]$, *is limited, the curve* L *is said to be* **rectifiable** *and the supremum* l *of the set* $\{\bar{l}(t_i)\}$ *is called the arc length of the curve* L.

Remark 1. It follows from the definition of the curve L represented parametrically and from the definition of the arc length l of such a curve that the length l is positive, $l > 0$.

Remark 2. There are curves which cannot be rectified. In the supplement to this chapter we shall give an example of a plane curve whose any part is not rectifiable.

In what follows we shall often use the following lemma.

Lemma. *Assume that* $\bar{l}^*(t_i)$ *is the length of a polygonal line inscribed into the curve* L *and corresponding to a partition* T^* *of the closed interval* $[\alpha, \beta]$, *and* $\bar{l}(t_i)$ *is the length of a polygonal line inscribed into the curve* L *and corresponding to partition* T *obtained from partition* T^* *by adding several new points. Then* $\bar{l}^*(t_i) \leqslant \bar{l}(t_i)$.

Proof. Evidently, it is sufficient to consider the case where one point γ is added to partition T^*. The polygonal line corresponding to partition T differs from that corresponding to partition T^* only by the fact that one element $M_{i-1}M_i$ is replaced by two elements $M_{i-1}C$ and CM_i (C is a point of the curve corresponding to the value γ of parameter t). Since the length of the side $M_{i-1}M_i$ of the triangle $M_{i-1}CM_i$ does not exceed the sum of the lengths of its other two sides* $M_{i-1}C$ and CM_i, we have $\bar{l}^*(t_i) \leqslant \bar{l}(t_i)$.

We will give here some properties of rectifiable curves.

1°. *If a curve* L *is rectifiable, then its arc length* l *does not depend on the parametrization of that curve.*

2°. *If a rectifiable curve* L *is divided, by means of a finite number of points* $M_0, M_1, \ldots, M_n$**, *into a finite number of curves* L_i, *then every one of these curves* L_i *is rectifiable and the sum of the lengths* l_i *of all the curves* L_i *is equal to the length* l *of the curve* L.

3°. *Suppose a curve* L *is specified parametrically by equations* (11.3). *We shall denote by* $l(t)$ *the arc length of the part* L_t *of the curve* L *whose points are specified by all the values of the parameter belonging to the closed interval* $[\alpha, t]$. *The function* $l(t)$ *is an increasing and continuous function of parameter* t. This function $l = l(t)$ will be called a *variable arc* on the curve L.

---

* This fact from geometry can be easily proved by an analytic method.

** In this case points $M_0, M_1, \ldots, M_n$ correspond to the values $t_0, t_1, \ldots, t_n$ of parameter t satisfying the conditions $\alpha = t_0 < t_1 < \ldots < t_n = \beta$.

4°. *A variable arc l can be chosen to be a parameter.* This parameter is known as a *natural parameter.*

The validity of property 4° follows immediately from property 3°. Indeed, since the variable arc $l = l(t)$ is an increasing and continuous function of parameter t, then parameter t can also be represented as a monotonic and continuous function $t = f(l)$ of the variable arc l and, therefore, the variable arc l can be chosen to be a parameter.

Proof of properties 1°-3°.

1°. Suppose we are given two parametrizations of the curve L, and t and s are the parameters of these parametrizations defined, respectively, on the closed intervals $[\alpha, \beta]$ and $[a, b]$. Since t is a strictly monotonic and continuous function of s, and s is a strictly monotonic and continuous function of t, each partition T of the interval $[\alpha, \beta]$ corresponds to a definite partition P of the interval $[a, b]$ and vice versa. It is evident that the polygonal lines inscribed into L and corresponding to the respective partitions of the intervals $[\alpha, \beta]$ and $[a, b]$ are identical and, therefore, their lengths $\overline{l}(t_i)$ and $\overline{l}(s_i)$ are equal. Consequently, the sets $\{\overline{l}(t_i)\}$ and $\{\overline{l}(s_i)\}$ are identical. From this it follows that the arc length of a curve does not depend on the parametrization of that curve.

2°. It is, evidently, sufficient to prove property 2° for the case where the curve L is divided by a point C into two curves, L_1 and L_2. Let us designate by γ the value of parameter t which is associated with point C. Then the points of the curve L_1 correspond to the values of parameter t belonging to the closed interval $[\alpha, \gamma]$ and the points of the curve L_2 correspond to the values of parameter t belonging to the interval $[\gamma, \beta]$. Let T_1 and T_2 be arbitrary partitions of the indicated intervals and T a partition of the interval $[\alpha, \beta]$ obtained by combining partitions T_1 and T_2. If $\overline{l}_1(t_i)$, $\overline{l}_2(t_i)$ and $\overline{l}(t_i)$ are the lengths of the polygonal lines inscribed into the curves L_1, L_2 and L and corresponding to partitions T_1, T_2 and T of the closed intervals indicated above, then, evidently,

$$\overline{l}_1(t_i) + \overline{l}_2(t_i) = \overline{l}(t_i). \tag{11.7}$$

The numbers $\overline{l}_1(t_i)$, $\overline{l}_2(t_i)$ and $\overline{l}(t_i)$ being positive, equality (11.7) and the rectifiability of the curve L imply that the sets $\{\overline{l}_1(t_i)\}$ and $\{\overline{l}_2(t_i)\}$ of the lengths of the open polygons inscribed into the curves L_1 and L_2, corresponding to various partitions of the closed intervals $[\alpha, \gamma]$ and $[\gamma, \beta]$ are bounded, that is, the curves L_1 and L_2 are rectifiable. It should be pointed out that equality (11.7) and the definition of the arc length of a curve imply that the lengths l_1, l_2 and l of the arcs of the curves L_1, L_2 and L satisfy the inequality*

$$l_1 + l_2 \leqslant l. \tag{11.8}$$

Assume $l_1 + l_2 < l$. Then the number

$$l - (l_1 + l_2) = \varepsilon \tag{11.9}$$

is positive. It follows from the definition of the arc length l of the curve L that for a positive number ε there is a partition T^* of the interval $[\alpha, \beta]$ such that the length $\overline{l}^*(t_i)$ of the polygonal line inscribed into the curve L and corresponding to that partition satisfies the inequality $l - \overline{l}^*(t_i) < \varepsilon$. Let us add a point γ

* It follows from equality (11.7) that for any partitions T_1 and T_2 of the closed intervals $[\alpha, \gamma]$ and $[\gamma, \beta]$ the inequality $\overline{l}_1(t_i) + \overline{l}_2(t_i) \leqslant l$ holds true. From this and from the definition of the supremum we obtain inequality (11.8).

to partition T^* and denote the resulting partition by T. Then, by virtue of the lemma presented in this subsection, the length $\bar{l}(t_i)$ of the polygonal line corresponding to partition T complies with the inequality $l - \bar{l}(t_i) < \varepsilon$. Since partition T of the interval $[\alpha, \beta]$ is obtained by combining some partitions T_1 and T_2 of the intervals $[\alpha, \gamma]$ and $[\gamma, \beta]$, the lengths $\bar{l}_1(t_i)$ and $\bar{l}_2(t_i)$ of the polygonal lines corresponding to those partitions satisfy relation (11.7). Therefore, there holds inequality $l - [\bar{l}_1(t_i) + \bar{l}_2(t_i)] < \varepsilon$. Since $\bar{l}_1(t_i) + \bar{l}_2(t_i) \leqslant l_1 + l_2$, all the more the inequality $l - (l_1 + l_2) < \varepsilon$ holds true. But this inequality contradicts equality (11.9), and so the assumption that $l_1 + l_2 < l$ is incorrect and, hence, by virtue of (11.8), $l_1 + l_2 = l$. We have thus established the validity of property 2°.

3°. It follows from property 2° and Remark 1 of this subsection that *the variable arc $l = l(t)$ is a strictly increasing positive function of parameter t.* To prove the continuity of the function $l(t)$, we shall make use of the following assertions.

(1) *Assume that ε is any fixed positive number, t is an arbitrary point belonging to the closed interval $[\alpha, \beta]$, and M is the respective point of the curve L. There exists a polygonal line, inscribed into the curve L, which has point M as its vertex and whose length differs from the length of the curve L by less than $\varepsilon/2$.*

(2) *The polygonal line can be so chosen that the length of its every element is less than $\varepsilon/2$.*

(3) *Let us choose the polygonal line as indicated in assertions* (1) *and* (2). *Then a part of the curve L subtended by any element of the open polygon being considered has a length less than ε.*

Let us verify that the formulated assertions and the monotonicity of the function $l(t)$ yield its continuity at any fixed point t of that interval (at points α and β the function $l(t)$ is continuous from the right and from the left, respectively).

We must prove that for any positive ε there is a positive δ such that the inequality $|l(t + \Delta t) - l(t)| < \varepsilon$ holds true at $|\Delta t| < \delta$.

Let us consider a partition T of the interval $[\alpha, \beta]$ associated with a polygonal line possessing the properties indicated in assertions (1) and (2). Let us denote by δ the minimum of the lengths of the two subintervals $[t_{k-1}, t_k]$, $[t_k, t_{k+1}]$ of partition T adjoining point $t = t_k$ of the interval $[\alpha, \beta]$. Assume that the increment Δt of the argument satisfies the condition $|\Delta t| < \delta$. For the sake of definiteness, we shall assume $\Delta t > 0$. Since $t < t + \Delta t < t + \delta \leqslant t_{k+1}$, there hold inequalities

$$l(t) < l(t + \Delta t) < l(t + \delta) \leqslant l(t_{k+1})$$

by virtue of a strict increase of the function $l(t)$. The inequality

$$l(t_{k+1}) - l(t) < \varepsilon$$

is valid by virtue of assertion 3. This fact and the preceding inequalities yield an inequality

$$l(t + \Delta t) - l(t) < \varepsilon$$

which holds for $0 < \Delta t < \delta$. The argument for the case $\Delta t < 0$ is the same.

We shall now prove assertions (1), (2) and (3).

Proof of assertion 1. Assume ε to be any fixed positive number. Since the length $l(\beta)$ of the whole curve L specified by parametric equations (11.3) is the supremum of the lengths of the open polygons inscribed into that curve and corresponding to various partitions of the interval $[\alpha, \beta]$, there is a partition T^{**} of the interval $[\alpha, \beta]$, for a given positive ε, for which the length of the corresponding polygonal line inscribed into the curve L differs from $l(\beta)$ by less than $\varepsilon/2$. Let us add a point t to partition T^{**}. By virtue of the lemma presented

in this subsection and the definition of the arc length, the length of the polygonal line corresponding to partition T^* of the interval $[\alpha, \beta]$ we have obtained differs from $l(\beta)$ by less than $\varepsilon/2$, and this polygonal line has as its vertex a point M of the curve corresponding to point t of the interval $[\alpha, \beta]$.

Proof of assertion 2. Since the functions $\varphi(t)$ and $\psi(t)$, continuous on the closed interval $[\alpha, \beta]$ are uniformly continuous on that interval, we can choose, proceeding from a given positive ε, a positive δ such that for any partition T of the interval $[\alpha, \beta]$ with the lengths of the subintervals $[t_{i-1}, t_i]$ less than δ the inequalities $|\varphi(t_i) - \varphi(t_{i-1})| < \frac{\varepsilon}{2\sqrt{2}}$, $|\psi(t_i) - \psi(t_{i-1})| < \frac{\varepsilon}{2\sqrt{2}}$ hold true. Inasmuch as the length $\overline{l}_i$ of the element of the polygonal line corresponding to the given partition is equal to $\sqrt{[\varphi(t_i) - \varphi(t_{i-1})]^2 + [\psi(t_i) - \psi(t_{i-1})]^2}$, $\overline{l}_i$ is, evidently, less than $\frac{\varepsilon}{2}$. Let us now consider any fixed partition T' of the closed interval $[\alpha, \beta]$ with the lengths of the subintervals less than δ and add to it the points of partition T^* (see the proof of assertion 1). As a result we shall obtain partition T which is associated with an open polygon inscribed into the curve L and satisfying all the conditions of assertion 2.

Proof of assertion 3. Suppose the polygonal line $M_0M_1 \ldots M_{k-1}M_kM_{k+1} \ldots M_n$ satisfies the conditions of assertions 1 and 2. Let us verify that the length of each part of the curve L subtended by any element of the polygonal line being considered is less than ε. In fact, assume that l_k is the length of the part $M_{k-1}M_k$ of the curve L and $\overline{l}_k$ is the length of the element $M_{k-1}M_k$ of the polygonal line. Then, by virtue of the conditions of assertion 1, the inequality $\sum_{k=1}^{n} (l_k - \overline{l}_k) < \frac{\varepsilon}{2}$ holds true. Since every summand $l_k - \overline{l}_k$ of the last sum is nonnegative, we have $l_k - \overline{l}_k < \frac{\varepsilon}{2}$. This inequality and the inequality $\overline{l}_k < \frac{\varepsilon}{2}$ yield the required inequality $l_k < \varepsilon$.

The concept of the arc length of a space curve specified by parametric equations (11.5) can be introduced by a complete analogy with the concept of the arc length of a plane curve. The lengths $l(t_i)$ of the polygonal lines inscribed into the curve L are considered, with, obviously,

$$\overline{l}(t_i) = \sum_{i=1}^{n} \sqrt{[\varphi(t_i) - \varphi(t_{i-1})]^2 + [\psi(t_i) - \psi(t_{i-1})]^2 + [\chi(t_i) - \chi(t_{i-1})]^2}.$$

The space curve L specified by equations (11.5) is said to be *rectifiable* if the set $\{\overline{l}(t_i)\}$ of the lengths of the polygonal lines inscribed into that curve is bounded. The least upper bound l of that set is called the arc length of the curve L.

It should be noted that rectifiable space curves possess properties 1°, 2°, 3° and 4° enumerated in this subsection. The proof of these properties can be carried out by a complete analogy with that for plane curves.

11.1.5. Sufficient conditions for rectifiability of a curve. Formulas for computing the arc length of a curve.

Theorem 11.1. *If the functions $x = \varphi(t)$ and $y = \psi(t)$ possess continuous derivatives on the closed interval $[\alpha, \beta]$, then the curve L specified by parametric equations* (11.3) *is rectifiable and its arc length l can be calculated by the formula*

$$l = \int_{\alpha}^{\beta} \sqrt{\varphi'^2(t) + \psi'^2(t)}\, dt. \tag{11.10}$$

Proof. Let us, first, prove that the curve L is rectifiable. For that purpose, let us transform expression (11.6) for the length $\bar{l}(t_i)$ of the polygonal line inscribed into the curve L and corresponding to the arbitrary partition T of the closed interval $[\alpha, \beta]$. Since the functions $\varphi(t)$ and $\psi(t)$ possess derivatives on the closed interval $[\alpha, \beta]$, we have, in accordance with Lagrange's formula, $\varphi(t_i) - \varphi(t_{i-1}) = \varphi'(\tau_i)\,\Delta t_i$, where $t_{i-1} < \tau_i < t_i$, $\Delta t_i = t_i - t_{i-1}$ and $\psi(t_i) - \psi(t_{i-1}) = \psi'(\tau_i^*)\,\Delta t_i$, where $t_{i-1} < \tau_i^* < t_i$. Substituting the expressions obtained for $\varphi(t_i) - \varphi(t_{i-1})$ and $\psi(t_i) - \psi(t_{i-1})$ into the right-hand side of expression (11.6), we obtain

$$\bar{l}(t_i) = \sum_{i=1}^{n} \sqrt{\varphi'^2(\tau_i) + \psi'^2(\tau_i^*)}\, \Delta t_i. \tag{11.11}$$

By the hypothesis, the functions $\varphi(t)$ and $\psi(t)$ possess continuous derivatives on the closed interval $[\alpha, \beta]$. Consequently, these derivatives are bounded and, therefore, there is a number M such that for all t belonging to the interval $[\alpha, \beta]$ the inequalities $|\varphi'(t)| \leqslant M$ and $|\psi'(t)| \leqslant M$ hold true. But then formula (11.11) yields

$$0 < \bar{l}(t_i) \leqslant \sum_{i=1}^{n} \sqrt{M^2 + M^2}\, \Delta t_i = M\sqrt{2} \sum_{i=1}^{n} \Delta t_i = M\sqrt{2}\,(\beta - \alpha).$$

Thus the set $\{l(t_i)\}$ of the lengths of the polygonal lines inscribed into the curve L and corresponding to various partitions T of the closed interval $[\alpha, \beta]$ is bounded, that is, the *curve L is rectifiable*. Let us denote by l the length of that curve and prove that the length l of the curve L can be calculated from formula (11.10). Note that the right-hand side of formula (11.11) resembles the integral sum

$$I\{t_i, \tau_i\} = \sum_{i=1}^{n} \sqrt{\varphi'^2(\tau_i) + \psi'^2(\tau_i)}\, \Delta t_i \tag{11.12}$$

of the integrable function $\sqrt{\varphi'^2(t) + \psi'^2(t)}$, the sum $I\{t_i, \tau_i\}$ corresponding to partition T of the interval $[\alpha, \beta]$ and to the given choice of points τ_i on the subintervals $[t_{i-1}, t_i]$ of that partition. *Let us prove that for any positive ε there is a positive δ such that for*

$\Delta < \delta$ ($\Delta = \max \Delta t_i$) *there holds an inequality*

$$|\bar{l}(t_i) - I| < \frac{\varepsilon}{2}, \tag{11.13}$$

where $I = \int_{\alpha}^{\beta} \sqrt{\varphi'^2(t) + \psi'^2(t)}\, dt$ is the limit of integral sums (11.12) for $\Delta \to 0$. In other words, *we shall prove that for sufficiently fine partitions T of the interval* $[\alpha, \beta]$ *the lengths* $\bar{l}(t_i)$ *of the open polygons inscribed into the curve L and corresponding to those partitions differ arbitrarily little from the integral I* appearing on the right-hand side of formula (11.10). We should note, first, that

$$\left|\sqrt{\varphi'^2(\tau_i) + \psi'^2(\tau_i^*)} - \sqrt{\varphi'^2(\tau_i) + \psi'^2(\tau_i)}\right| \leqslant$$
$$\leqslant |\psi'(\tau_i^*) - \psi'(\tau_i)| \leqslant M_i - m_i\,^*, \tag{11.14}$$

where M_i and m_i are the supremum and the infimum of the function $\psi'(t)$ on the subinterval $[t_{i-1}, t_i]$. In accordance with (11.11), (11.12) and (11.14), there hold inequalities

$$|\bar{l}(t_i) - I\{t_i, \tau_i\}| = \left|\sum_{i=1}^{n} \left(\sqrt{\varphi'^2(\tau_i) + \psi'^2(\tau_i^*)} - \right.\right.$$
$$\left.\left. - \sqrt{\varphi'^2(\tau_i) + \psi'^2(\tau_i)}\right) \Delta t_i\right| \leqslant \sum_{i=1}^{n} \left|\sqrt{\varphi'^2(\tau_i) + \psi'^2(\tau_i^*)} - \right.$$
$$\left. - \sqrt{\varphi'^2(\tau_i) + \psi'^2(\tau_i)}\right| \Delta t_i \leqslant \sum_{i=1}^{n} (M_i - m_i)\, \Delta t_i = S - s, \tag{11.15}$$

where S and s are the upper and the lower sum of the function $\psi'(t)$ for partitioning the closed interval $[\alpha, \beta]$. The functions $\sqrt{\varphi'^2(t) + \psi'^2(t)}$ and $\psi'(t)$ being integrable on the interval $[\alpha, \beta]$ (this follows from the continuity of the derivatives $\varphi'(t)$ and $\psi'(t)$ on the interval $[\alpha, \beta]$), it follows from the definition of integrability and from Theorem 10.1 (see 10.1 and 10.3) that for any

---

* To obtain inequalities (11.14), we have made use of the inequality $\left|\sqrt{a^2 + b^{*2}} - \sqrt{a^2 + b^2}\right| \leqslant |b^* - b|$, where $a^2 = \varphi'^2(\tau_i)$, $b^{*2} = \psi'^2(\tau_i^*)$ and $b^2 = \psi'^2(\tau_i)$, and the inequality $|\psi'(\tau_i^*) - \psi'(\tau_i)| \leqslant M_i - m_i$.

The second inequality is obvious since the difference between any values of a function does not exceed the difference between its supremum and infimum. Let us prove the first of the indicated inequalities. We have

$$|\sqrt{a^2 + b^{*2}} - \sqrt{a^2 + b^2}| = \frac{|b^{*2} - b^2|}{\sqrt{a^2 + b^{*2}} + \sqrt{a^2 + b^2}} \leqslant$$
$$\leqslant \frac{|b^* - b|\,|b^* + b|}{\sqrt{b^{*2}} + \sqrt{b^2}} \leqslant \frac{|b^* - b|\,(|b^*| + |b|)}{|b^*| + |b|} = |b^* - b|.$$

$\varepsilon > 0$, there is a positive δ such that at $\Delta < \delta$ ($\Delta = \max \Delta = t_i$) there hold inequalities

$$|I\{t_i, \tau_i\} - I| < \frac{\varepsilon}{4} \text{ and } S - s < \frac{\varepsilon}{4}. \tag{11.16}$$

Therefore, for $\Delta < \delta$, by virtue of (11.15) and (11.16), the inequalities $|\bar{l}(t_i) - I| = |\bar{l}(t_i) - I\{t_i, \tau_i\} + I\{t_i, \tau_i\} - I| \leqslant |\bar{l}(t_i) - I\{t_i, \tau_i\}| + |I\{t_i, \tau_i\} - I| < \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}$ hold true. Thus we have proved the validity of inequality (11.13).

Let us prove now that among various polygonal lines whose lengths $\bar{l}(t_i)$ *satisfy inequality* (11.13), *there are polygonal lines whose lengths differ from the arc length* l *of the curve* L *by less than* $\varepsilon/2$.

Since l is the least upper bound of the set $\{\bar{l}(t_i)\}$ of the lengths of the polygonal lines inscribed into the curve L and corresponding to various partitions of the closed interval $[\alpha, \beta]$, there is a partition T^* of that interval such that the length $\bar{l}^*(t_i)$ of the respective open polygon satisfies the inequalities

$$0 \leqslant I - \bar{l}^*(t_i) < \frac{\varepsilon}{2}. \tag{11.17}$$

Let us now divide each of the subintervals $[t_{i-1}, t_i]$ of partition T^* into parts so fine that the maximum length Δ of partition T of the closed interval $[\alpha, \beta]$, obtained by combining the indicated partitions, is less than δ, $\Delta < \delta$. It is evident that the length $\bar{l}(t_i)$ of the polygonal line corresponding to partition T satisfies inequality (11.13). Since the vertices of the open polygon corresponding to partition T^* are also the vertices of the open polygon corresponding to partition T, the length $\bar{l}(t_i)$ satisfies inequalities $0 < \bar{l}^*(t_i) \leqslant \bar{l}(t_i) \leqslant l$ in accordance with the lemma presented in this subsection and, therefore, by virtue of inequality (11.17), there holds an inequality

$$0 \leqslant l - \bar{l}(t_i) < \frac{\varepsilon}{2}. \tag{11.18}$$

Thus we have proved that among the polygonal lines whose lengths $l(t_i)$ satisfy inequality (11.13) there are those whose lengths $\bar{l}(t_i)$ satisfy inequality (11.18). Comparing inequalities (11.13) and (11.18), we obtain the following inequality:

$$|l - I| < \varepsilon.$$

The number ε being arbitrary, it follows that $l = I$, and that is the proof of the theorem.

Remark 1. *If the functions* $\varphi(t)$ *and* $\psi(t)$ *possess bounded derivatives on the closed interval* $[\alpha, \beta]$, *then the curve* L *specified by equations* (11.1) *is rectifiable.* Indeed, while proving Theorem 11.1 we have established that under the condition of boundedness of the derivatives of the functions $\varphi(t)$ and $\psi(t)$ the lengths $l(t_i)$ of the polygonal lines inscribed into the curve L and corresponding to various partitions T of the interval $[\alpha, \beta]$ are bounded.

Remark 2. *Formula* (11.10) *for calculating the arc length is valid if the derivatives* $\varphi'(t)$ *and* $\psi'(t)$ *are defined and integrable on the closed interval* $[\alpha, \beta]$. In fact, the integrability of these derivatives implies their boundedness and, therefore, by virtue of Remark 1, rectifiability of the curve L. Let us also note that to derive inequalities (11.14), (11.15) and (11.16), and, consequently, inequality (11.13), it is sufficient that the derivatives $\varphi'(t)$ and $\psi'(t)$ exist and be integrable, since it yields, in accordance with Supplement 1 to Chap. 10, integrability of the function $\sqrt{\varphi'^2(t) + \psi'^2(t)}$. The rest of the reasoning is the same as that used in the proof of Theorem 11.1.

Remark 3. *If the curve* L *is the graph of the function* $y = f(x)$, *possessing a continuous derivative* $f'(x)$ *on the closed interval* $[a, b]$ *then the curve* L *is rectifiable and the length* l *of the arc* L *can be found from the formula*

$$l = \int_a^b \sqrt{1 + f'^2(x)}\, dx. \tag{11.19}$$

To prove this statement, let us pay attention to the fact that the graph of the function being discussed is a curve specified by the parametric equations $x = t$, $y = f(t)$, $a \leqslant t \leqslant b$, all the conditions of Theorem 11.1 being evidently satisfied. For that reason, putting $\varphi(t) = t$, $\psi(t) = f(t)$ in formula (11.10) and replacing the variable of integration t by x, we shall obtain formula (11.19). We shall also take into consideration the fact that if the curve L is specified by the polar equation $r = r(\theta)$, $\theta_1 \leqslant \theta \leqslant \theta_2$ and the function $r(\theta)$ possesses a continuous derivative on the closed interval $[\theta_1, \theta_2]$, then the curve L is rectifiable and the length l of the arc L can be found from the formula

$$l = \int_{\theta_1}^{\theta_2} \sqrt{r^2(\theta) + r'^2(\theta)}\, d\theta. \tag{11.20}$$

To prove this statement, let us make use of formulas of transition from the polar to the Cartesian coordinates:

$$x = r(\theta)\cos\theta, \quad y = r(\theta)\sin\theta.$$

Thus we see that the curve L is specified by parametric equations, with the functions $\varphi = r(\theta)\cos\theta$ and $\psi = r(\theta)\sin\theta$ satisfying the conditions of Theorem 11.1. Substituting the indicated values of φ and ψ into (11.10), we obtain formula (11.20).

Let us formulate the sufficient conditions for rectifiability of a space curve.

If the functions $\varphi(t)$, $\psi(t)$ *and* $\chi(t)$ *possess continuous derivatives on the closed interval* $[\alpha, \beta]$, *then the curve* L *specified by equations* (11.5) *is rectifiable and the length* l *of its arc can be found from the formula*

$$l = \int_{\alpha}^{\beta} \sqrt{\varphi'^2(t) + \psi'^2(t) + \chi'^2(t)}\, dt. \tag{11.21}$$

The proof is the same as that of Theorem 11.1.

Remark 4. If the functions $\varphi(t)$, $\psi(t)$ and $\chi(t)$ possess bounded derivatives on the closed interval $[\alpha, \beta]$, then the curve L specified by equations (11.5) is rectifiable. If in that case the derivatives of the indicated functions are integrable on the interval $[\alpha, \beta]$, then the length l of the arc L can be calculated by formula (11.21) (see Remarks 1 and 2).

11.1.6. Differential of an arc. Suppose the function $x = \varphi(t)$ and $y = \psi(t)$ possess continuous derivatives on the closed interval $[\alpha, \beta]$. In that case, by virtue of Theorem 11.1, the variable arc $l(t)$ can be represented by the following formula:

$$l(t) = \int_{\alpha}^{t} \sqrt{\varphi'^2(\tau) + \psi'^2(\tau)}\, d\tau. \tag{11.22}$$

Since the integrand function on the right-hand side of formula (11.22) is continuous, the function $l(t)$ is differentiable, with

$$l'(t) = \sqrt{\varphi'^2(t) + \psi'^2(t)}$$

(see 10.7.1). Squaring both sides of the last equality and then multiplying by dt^2, we obtain

$$[l'(t)\, dt]^2 = [\varphi'(t)\, dt]^2 + [\psi'(t)\, dt]^2. \tag{11.23}$$

Since $l'(t)\, dt = dl$, $\varphi'(t)\, dt = dx$, $\psi'(t)\, dt = dy$, it follows from (11.23) that

$$dl^2 = dx^2 + dy^2. \tag{11.24}$$

From formula (11.24) it follows, in particular, that if the arc l is taken as the parameter, that is, if $x = g(l)$ and $y = h(l)$, then

$$\left(\frac{dx}{dl}\right)^2 + \left(\frac{dy}{dl}\right)^2 = 1. \tag{11.25}$$

It should be noted that under the condition of continuity of the derivatives of the functions $x = \varphi(t)$, $y = \psi(t)$ and $z = \chi(t)$, the formula

$$dl^2 = dx^2 + dy^2 + dz^2 \tag{11.26}$$

is valid for the differential dl of the arc of the space curve specified by parametric equations (11.5). From formula (11.26) it follows that

if the variable arc l is taken as the parameter, then

$$\left(\frac{dx}{dl}\right)^2+\left(\frac{dy}{dl}\right)^2+\left(\frac{dz}{dl}\right)^2=1. \tag{11.27}$$

11.1.7. Some examples of arc length computation. 1°. The arc length of the cycloid* $x = a(t - \sin t)$, $y = a(1 - \cos t)$, $0 \leqslant t \leqslant 2\pi$. In the case being considered $\varphi' = a(1 - \cos t)$, $\psi' = a \sin t$. Therefore, in accordance with formula (11.10),

$$l = \int_0^{2\pi} a \sqrt{(1-\cos t)^2 + \sin^2 t}\, dt =$$

$$= 2a \int_0^{2\pi} \sin \frac{t}{2}\, dt = -4a \cos \frac{t}{2}\Big|_0^{2\pi} = 8a.$$

2°. A catenary line is the graph of the function $y = a \cosh \frac{x}{a}$. Let us find the length of the portion of the catenary corresponding to the closed interval $[0, x]$. We have, in accordance with formula (11.19),

$$l(x) = \int_0^x \sqrt{1+y'^2(\xi)}\, d\xi = \int_0^x \sqrt{1+\sinh^2\frac{\xi}{a}}\, d\xi =$$

$$= \int_0^x \cosh \frac{\xi}{a}\, d\xi = a \sinh \frac{x}{a}.$$

3°. Let us find the variable arc of the ellipse $\frac{x^2}{a^2}+\frac{y^2}{b^2}=1$, $a > b$, reckoned from point $M_0(0, b)$. Let us consider the parametric equations of an ellipse $x = a \sin t$, $y = b \cos t$, $0 \leqslant t \leqslant 2\pi$. By formula (11.22) we have

$$l(t) = \int_0^t \sqrt{\varphi'^2(\tau)+\psi'^2(\tau)}\, d\tau = \int_0^t \sqrt{a^2\cos^2\tau + b^2\sin^2\tau}\, d\tau =$$

$$= a \int_0^t \sqrt{1-e^2\sin^2\tau}\, d\tau = aE(e, t).$$

The number $e = \frac{\sqrt{a^2-b^2}}{a}$ is called an eccentricity of an ellipse. The indefinite integral $\int \sqrt{1-e^2\sin^2 t}\, dt$ vanishing at $t=0$ is called an elliptic integral of the 2nd kind and denoted $E(e, t)$ (see 7.11).

* A cycloid is a plane curve described by a point of a circle of radius a rolling without slipping along a straight line.

11.2. AREA OF A PLANE FIGURE *

11.2.1. The concept of squareability of a plane figure. Area of a squareable plane figure. We are familiar with the concept of an area of a plane figure bounded by a polygon** from the course of elementary mathematics. Here we shall introduce the concept of the area of a *plane figure* Q which is a portion of a plane bounded by a simple closed curve L***. In this case, the curve L will be called a boundary of the figure Q.

A polygon is said to be *inscribed* into the figure Q if every point of that polygon belongs to the figure Q or to its boundary. If all the points of a plane figure and of its boundary belong to a certain polygon, we shall say that the indicated polygon is *circumscribed* about the figure Q.

It is clear that the area of any polygon inscribed into the figure Q does not exceed the area of any polygon circumscribed about that figure Q.

Suppose $\{S_i\}$ is a number set of the areas of polygons inscribed into a plane figure Q and $\{S_d\}$ is a number set of the areas of polygons circumscribed about that figure Q. It is evident that the set $\{S_i\}$ is bounded above (by the area of any polygon circumscribed about figure Q) and the set $\{S_d\}$ is bounded below (say, by number zero). Let us designate by $\underline{P}$ the least upper bound of the set $\{S_i\}$ and by $\overline{P}$ the greatest lower bound of the set $\{S_d\}$. *The numbers* $\underline{P}$ *and* $\overline{P}$ *are called, respectively, the lower area and the upper area of the figure* Q. Note that the lower area $\underline{P}$ of the figure Q does not exceed the upper area $\overline{P}$ of that figure, that is, $\underline{P} \leqslant \overline{P}$. Indeed, let us assume that the contrary is true, i.e. $\underline{P} > \overline{P}$. Then, assuming $\frac{\underline{P}-\overline{P}}{2} = \varepsilon > 0$ and taking into account the definition of the supremum and the infimum, we shall find a polygon inscribed into the figure Q such that the area S_i of the polygon will be greater than the number $\underline{P} - \varepsilon = \frac{\underline{P}+\overline{P}}{2}$, i.e. $\frac{\underline{P}+\overline{P}}{2} < S_i$, and a polygon circumscribed about the figure Q such that its area S_d will be smaller than the number

---

* The reader will often come across the concepts of an area of a plane figure and of an arbitrary set of points of a plane in the second volume of this course.

** We shall use the term polygon for a part of a plane bounded by a simple closed broken line.

*** Note that a simple closed plane curve L divides the area into two parts, interior and exterior. This assertion was proved by Camille Jordan, a French mathematician (1838-1922).

$\overline{P}+\varepsilon=\frac{P+\overline{P}}{2}$, i.e. $S_d<\frac{P+\overline{P}}{2}$. Comparing the two inequalities we have just obtained we shall find that $S_d<S_i$. But this inequality cannot exist since the area S_d of any circumscribed polygon *is not smaller* than the area S_i of any inscribed polygon.

We wish now to introduce the notion of squareability of a plane figure.

Definition. *A plane figure Q is said to be squareable if the upper area $\overline{P}$ of that figure coincides with its lower area $\underline{P}$. In that case, the number $P=\underline{P}=\overline{P}$ is called the area of the figure Q.*

Remark. An example of a nonsquareable figure will be given in the supplement to this chapter.

The following theorem holds true.

Theorem 11.2. *For the plane figure Q to be squareable it is necessary and sufficient that for any positive number ε there should be a polygon circumscribed about the figure Q and a polygon inscribed into the figure Q such that the difference S_d-S_i in the areas of these polygons is smaller than ε, $S_d-S_i<\varepsilon$.*

Proof. (1) **Necessity.** Suppose the figure Q is squareable, i.e. $\underline{P}=\overline{P}=P$. Since $\underline{P}$ and $\overline{P}$ are the supremum and the infimum of the sets $\{S_i\}$ and $\{S_d\}$, there is a polygon inscribed into the figure Q, for any $\varepsilon>0$, such that the polygon area S_i differs from $\underline{P}=P$ by less than $\varepsilon/2$, i.e. $P-S_i<\varepsilon/2$. For the same $\varepsilon>0$, we can indicate a circumscribed polygon whose area S_d differs from $\overline{P}=P$ by less than $\varepsilon/2$, i.e. $S_d-P<\varepsilon/2$. Adding up the inequalities obtained, we shall find that $S_d-S_i<\varepsilon$.

2. **Sufficiency.** Assume that S_d and S_i are the areas of the polygons for which $S_d-S_i<\varepsilon$. Since $S_i\leqslant\underline{P}\leqslant\overline{P}\leqslant S_d$, we have $\overline{P}-\underline{P}<\varepsilon$. The number ε being arbitrary, it follows that $\underline{P}=\overline{P}$. Thus we see that the figure is squareable. We have proved the theorem.

We shall say that *the boundary of a plane figure Q has an area equal to zero*, if for any positive number ε there is a polygon circumscribed about the figure Q and a polygon inscribed into the figure Q such that the difference S_d-S_i of their areas is less than ε. Theorem 11.2 can, evidently, be formulated as follows.

For the plane figure Q to be squareable it is necessary and sufficient that the area of its boundary be equal to zero.

Remark. In all our reasoning, we could have considered an arbitrary set of points of a plane instead of a plane figure.

Let us establish the sufficient condition for squareability of a plane figure.

Theorem 11.3. *If the boundary L of the plane figure Q is a rectifiable curve, then the figure Q is squareable.*

Proof. Suppose l^* is the length of the curve L. We shall assume that the curve L has been parametrized with the aid of a natural parameter l, $0 \leqslant l \leqslant l^*$. In this case, the curve L being closed, its boundary points corresponding to the values 0 and l^* of the parameter l coincide. Assume ε to be an arbitrary positive number. Let us divide the closed interval $[0, l^*]$ by the points $0 = l_0 < l_1 < \ldots < l_n = l^*$ into n equal parts of length smaller than $\varepsilon/9l^*$. We shall consider the polygonal line $M_0M_1, \ldots, M_n$ $(M_0 = M_n)$ inscribed into the curve L and corresponding to the indicated partitioning of the interval $[0, l^*]$. The points $M_0, M_1, \ldots, M_n$ divide the curve L into the parts $L_1, L_2, \ldots, L_n$ whose lengths are equal to $\frac{l^*}{n} < \frac{\varepsilon}{9l^*}$. It is evident that the lengths of the elements $M_{i-1}M_i$ of the indicated polygonal line $M_0M_1 \ldots M_n$ do not exceed l^*/n. Let us insert each element $M_{i-1}M_i$ inside a square with side $3l^*/n$ as shown in Fig. 11.4. It is easy to verify that the arc L_i subtended by the element $M_{i-1}M_i$ is located inside this square since the distance from any point lying outside the square or on its boundary to every point M_{i-1}, and M_i is not smaller than l^*/n and, therefore, if some point M of the arc L_i was outside the square or on its boundary, the polygonal line $M_{i-1}MM_i$ inscribed into that arc would have the length not smaller than $2l^*/n$, i.e. larger than the length l^*/n of the arc L_i which is impossible. The combination of all these squares constructed on all the elements of the polygonal line $M_0M_1 \ldots M_n$ is a polygon containing the curve L inside, the boundary of this polygon constituting, evidently, the combination of the boundaries of the polygon inscribed into the figure Q and of the polygon circumscribed about Q. It is also evident that the difference $S_d - S_i$ in the areas of these polygons is equal to the area of the indicated figure and the area of this figure does not exceed the sum S of the areas of the squares described earlier. Since $S = n\frac{9l^{*2}}{n^2} = 9l^*\frac{l^*}{n} < \varepsilon$ (the last inequality follows from the fact that $\frac{l^*}{n} < \frac{\varepsilon}{9l^*}$), we have $S_d - S_i < \varepsilon$. Therefore, in accordance with Theorem 11.2 the figure Q is squareable. We have proved the theorem.

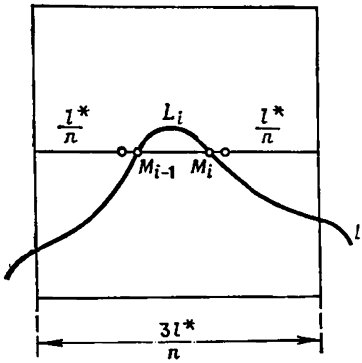


Fig. 11.4

11.2.2. Area of a curvilinear trapezoid. *A curvilinear trapezoid* is a figure bounded by the graph of a continuous and nonnegative function $f(x)$ defined on the closed interval $[a, b]$, by the ordinates drawn at points a and b, and by a segment of the x-axis between points a and b (Fig. 11.5). Let us prove the following assertion.

A curvilinear trapezoid is a squareable figure whose area P can be calculated by the formula

$$P = \int_a^b f(x)\,dx. \tag{11.28}$$

Proof. Since the function continuous on the closed interval $[a, b]$ is integrable, there can be indicated a partition T of the interval $[a, b]$, for any positive number ε, such that the difference $S - s < \varepsilon$,

where S and s are, respectively, the upper and the lower sum of partition T. But S and s are equal, respectively, to S_d and S_i, where S_d and S_i are the areas of the stepped figures (polygons), the first of which contains a curvilinear trapezoid and the second is contained in a curvilinear trapezoid (Figure 11.5 shows the indicated stepped

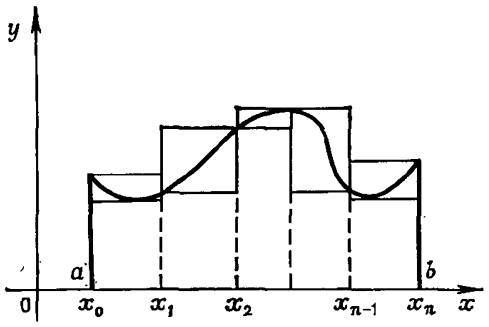


Fig. 11.5

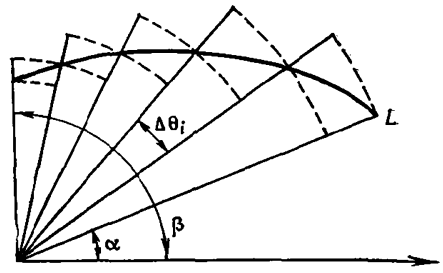


Fig. 11.6

figures as well). Since $S_d - S_i < \varepsilon$, the curvilinear trapezoid is squareable by virtue of Theorem 11.2. Since for $\Delta \to 0$ the limit of the upper and lower sums is equal to $\int_a^b f(x)\,dx$ and $s \leqslant P \leqslant S$, the area P of the curvilinear trapezoid can be found from formula (11.28).

Remark. If the function $f(x)$ is continuous and nonpositive on the closed interval $[a, b]$, the value of the integral $\int_a^b f(x)\,dx$ is equal to the area of a curvilinear trapezoid, taken with a negative sign and bounded by the graph of the function $f(x)$, the ordinates at points a and b and a segment of the x-axis between points a and b. Therefore, if $f(x)$ changes sign, then the integral $\int_a^b f(x)\,dx$ is equal to the sum of the areas of the curvilinear trapezoids, taken with a definite sign, located above or below the x-axis, the areas of the first trapezoids being taken with a plus sign and the areas of the second with a minus sign.

11.2.3. Area of a curvilinear sector. Suppose the curve L is expressed in polar coordinates by means of the equation $r = r(\theta)$, $\alpha \leqslant \theta \leqslant \beta$ (Fig. 11.6), with the function $r(\theta)$ being continuous and nonnegative on the closed interval $[\alpha, \beta]$. The plane figure bounded by the curve L and by two rays forming angles α and β with the polar axis will be called a *curvilinear sector.*

Let us prove the following assertion. *A curvilinear sector is a squareable figure whose area P can be calculated by the formula*

$$P=\frac{1}{2}\int_{\alpha}^{\beta} r^2(\theta)\,d\theta. \tag{11.29}$$

Proof. Let us consider a partition T of the closed interval $[\alpha, \beta]$ by points $\alpha=\theta_0<\theta_1<\ldots<\theta_n=\beta$, and for each subinterval $[\theta_{i-1}, \theta_i]$ let us construct circular sectors whose radii are equal to the minimum r_i and the maximum R_i values of $r(\theta)$ on the interval $[\theta_{i-1}, \theta_i]$. As a result, we shall obtain two fan-like figures the first of which is contained in the curvilinear sector and the second contains the curvilinear sector inside (these fan-like figures are shown in Fig. 11.6). The areas $\overline{S}_i$ and $\overline{S}_d$ of the indicated fan-like figures are equal, respectively, to $\frac{1}{2}\sum_{i=1}^{n} r_i^2\Delta\theta_i$ and $\frac{1}{2}\sum_{i=1}^{n} R_i^2\Delta\theta_i$. Note that the first of these sums is the lower sum s for the function $\frac{1}{2}r^2(\theta)$ for the indicated partition T of the interval $[\alpha, \beta]$ and the second is the upper sum S for the same function and the same partition. The function $\frac{1}{2}r^2(\theta)$ being integrable on the interval $[\alpha, \beta]$, the difference $S-s=\overline{S}_d-\overline{S}_i$ can be arbitrarily small. For instance, for any fixed positive ε this difference can be made smaller than $\varepsilon/2$. Let us now inscribe into the interior fan-like figure a polygon Q_i with area S_i for which $\overline{S}_i-S_i<\frac{\varepsilon}{4}$, and circumscribe about the exterior fan-like figure a polygon Q_d with area S_d for which $S_d-\overline{S}_d<\frac{\varepsilon}{4}$*. It is evident that the first polygon is inscribed into the curvilinear sector and the second is circumscribed about it. Since the inequalities

$$S_i<\overline{S}_i\leqslant\frac{1}{2}\int_{\alpha}^{\beta} r^2(\theta)\,d\theta<\overline{S}_d<S_d \tag{11.30}$$

are valid, we evidently have $S_d-S_i<\varepsilon$. The number ε being arbitrary, the last inequality yields the squareability of the curvilinear sector. Inequalities (11.30) yield the validity of formula (11.29).

* The fan-like figures being considered consist of circular sectors. Each sector is squareable and, consequently, the fan-like figures are squareable too. Therefore, polygons can be found for these figures whose areas S_i and S_d satisfy the indicated inequalities.

11.2.4. Examples of computing areas. 1°. Find the area P of the figure Q bounded by the graphs of the functions $y = x^\alpha$ and $x = y^\alpha$, $\alpha \geqslant 1$ (Fig. 11.7). The figure Q being symmetrical with respect to the bisector of the first coordinate angle, its area can be obtained by means of subtracting from unity (the area of the square) the

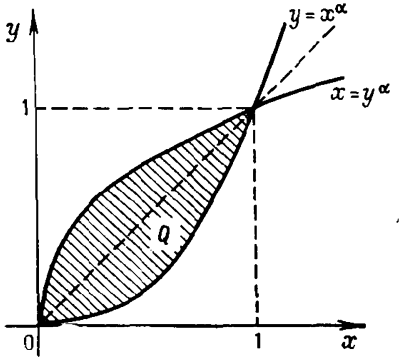


Fig. 11.7

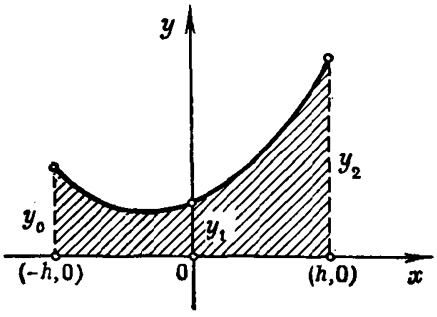


Fig. 11.8

doubled area of a curvilinear trapezoid determined by the graph of the function $y = x^\alpha$ on the closed interval [0, 1]. Thus, from formula (11.28) we have

$$P = 1 - 2\int_0^1 x^\alpha\,dx = 1 - 2\left[\frac{x^{\alpha+1}}{\alpha+1}\right]\Big|_0^1 = 1 - \frac{2}{\alpha+1} = \frac{\alpha-1}{\alpha+1}.$$

2°. Through three points with coordinates $(-h, y_0)$, $(0, y_1)$, (h, y_2) we can draw only one parabola $y = Ax^2 + Bx + D$ (or a straight line if the points lie on one straight line). Indeed, a system of equations with respect to A, B, D*

$$\left.\begin{aligned} Ah^2 - Bh + D &= y_0, \\ D &= y_1, \\ Ah^2 + Bh + D &= y_2 \end{aligned}\right\}$$

has a unique solution, namely,

$$A = \frac{y_0 - 2y_1 + y_2}{2h^2}, \quad B = \frac{y_2 - y_0}{2h}, \quad D = y_1.$$

Let us express the area P of the curvilinear trapezoid determined by the indicated parabola, the ordinates at points $(-h, 0)$ and $(h, 0)$ and the segment of the x-axis connecting these points (Fig. 11.8)

* These equations express the conditions of the positioning of the points $(-h, y_0)$ $(0, y_1)$ and (h, y_2) on the parabola $y = Ax^2 + Bx + D$.

using the ordinates y_0, y_1 and y_2. Since by formula (11.28) we have

$$P = \int_{-h}^{h} (Ax^2 + Bx + D)\,dx = \left[\frac{Ax^3}{3} + \frac{Bx^2}{2} + Dx\right]\Big|_{-h}^{h} =$$

$$= \frac{2Ah^3}{3} + 2Dh,$$

we shall find, taking into account the expressions for A and D, that

$$P = \frac{h}{3}(y_0 + 4y_1 + y_2).$$

3°. Find the area P of a trifolium $r = a\cos 3\theta$ (Fig. 11.9). It is clear from the drawing that the whole area of the trifolium is equal to the area of the hatched part of the trifolium increased six times which corresponds to the variation of θ from 0 to $\pi/6$. Therefore, from formula (11.29), we find

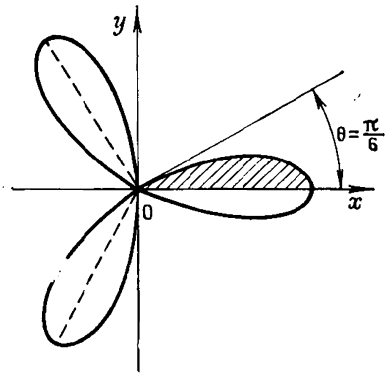


Fig. 11.9

$$P = 6\frac{a^2}{2}\int_0^{\pi/6} \cos^2 3\theta\,d\theta = \frac{\pi a^2}{4}.$$

11.3. VOLUMES OF BODIES AND AREAS OF SURFACES

11.3.1. The concept of cubability and volume. Suppose E is some finite body*. Let us consider various polygons inscribed into body E and various polygons circumscribed about it. Computation of the volume of the polygon reduces to the computation of the volumes of tetrahedra (triangular pyramids). Therefore, we shall consider the concept of the volume of a polygon to be familiar to us.

Assume that $\{V_i\}$ is the numerical set of the volumes of the polygons inscribed into body E, and $\{V_d\}$ is the numerical set of the volumes of the polygons circumscribed about E. The set $\{V_i\}$ is bounded above (by the volume of any circumscribed polygon) and the set $\{V_d\}$ is bounded below (say, by number zero). Let us denote by $\underline{V}$ the least upper bound of the set $\{V_i\}$ and by $\overline{V}$ the greatest lower bound of the set $\{V_d\}$. The numbers $\underline{V}$ and $\overline{V}$ are called, respectively, the *lower volume* and the *upper volume* of body E.

* We shall use the term body for a part of space bounded by a closed nonintersecting surface.

Note that the lower volume $\underline{V}$ of body E does not exceed its upper volume $\overline{V}$, i.e. $\underline{V} \leqslant \overline{V}$. To verify this fact, it is sufficient to use the same arguments as in the proof of the inequality $\underline{P} \leqslant \overline{P}$ (see 11.2.1).

We wish now to introduce the concept of the cubability of a body.

Definition. *A body E is said to be* **cubable** *if the upper volume $\overline{V}$ of that body coincides with its lower volume $\underline{V}$. In this case the number $V = \underline{V} = \overline{V}$ is called the volume of body E.*

The following theorem is valid.

Theorem 11.4. *For a body E to be cubable, it is necessary and sufficient that for any positive number ε there should be a polygon circumscribed about body E and a polygon inscribed into it such that the difference $V_d - V_i$ of their volumes be less than ε.* The proof of this theorem can be carried out by complete analogy with that of Theorem 11.2 (see 11.2.1).

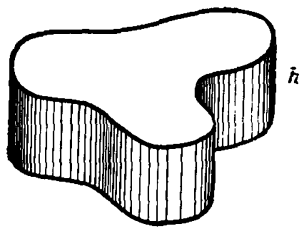


Fig. 11.10

11.3.2. Cubability of some classes of bodies. We shall use the term *cylinder* for a body bounded by a cylindrical surface with its elements being parallel to some axis and by two planes perpendicular to that axis. When these planes intersect a cylindrical surface, they generate plane figures called the *bases* of the cylinder, and the distance h between the bases of the cylinder is called the *altitude* of the cylinder (Fig. 11.10).

Let us prove the following assertion. *If the base of the cylinder E is a squareable figure Q, then the cylinder is a squareable body, the volume V of the cylinder E being equal to Ph, where P is the area of the base Q and h is the altitude of the cylinder.*

Since the figure Q is squareable, we can choose, for any positive number ε, polygons inscribed into the figure and circumscribed about it, such that the difference $S_d - S_i$ of their areas is less than ε/h. The volumes V_d and V_i of the prisms with altitude h whose bases are the indicated polygons, are equal, respectively, to $S_d h$ and $S_i h$. Therefore, $V_d - V_i = (S_d - S_i)\, h < \frac{\varepsilon}{h}\, h = \varepsilon$. Inasmuch as these prisms are, respectively, the polygons circumscribed about body E and inscribed into it, body E is cubable in accordance with Theorem 11.4. Since $V_i \leqslant Ph \leqslant V_d$, the volume of the cylinder is equal to Ph.

The assertion we have proved yields the *cubability of stepped bodies* (a stepped body is the combination of a finite number of cylinders positioned so that the upper base of each preceding cyl-

inder is in the same plane as the lower base of the succeeding cylinder, see Fig. 11.11).

Remark. The following obvious assertion is valid. *If for any positive number* ε *there is a stepped figure circumscribed about body* E *and a stepped figure inscribed into it such that the difference* $V_d - V_i$ *of their volumes is smaller than* ε, *then body* E *is cubable.*

Let us use this remark to prove the *cubability of solids of revolution* (Fig. 11.12). In fact, let us prove the following assertion.

Assume the function $y = f(x)$ *to be continuous on the closed interval* $[a, b]$. *Then a body* E *generated by the rotation about the x-axis of*

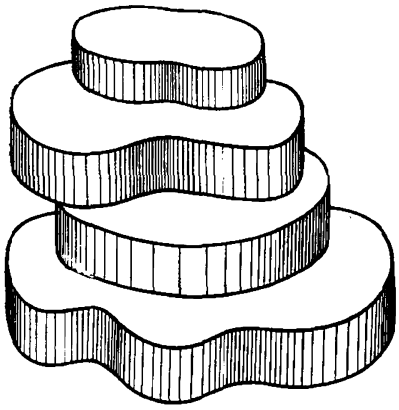


Fig. 11.11

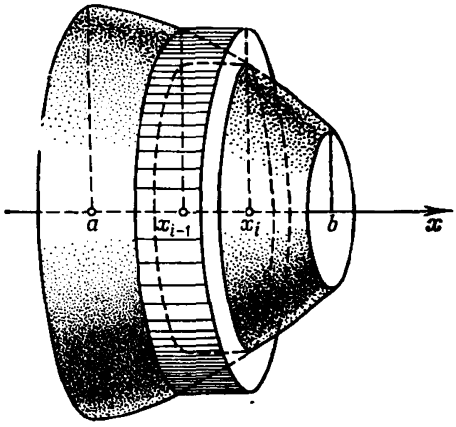


Fig. 11.12

a curvilinear trapezoid bounded by the graph of the function $f(x)$, *by the ordinates at points* a *and* b *and by the segment of the Ox axis connecting* a *and* b, *is cubable and its volume* V *can be found from the formula*

$$V = \pi \int_a^b f^2(x)\, dx. \tag{11.31}$$

Proof. Assume that T is a partition of the closed interval $[a, b]$ by points $a = x_0 < x_1 < \ldots < x_n = b$, m_i and M_i are the infimum and the supremum of $f(x)$ on the subinterval $[x_{i-1}, x_i]$. Let us construct two rectangles with altitudes m_i and M_i on each interval of this kind (Fig. 11.12 shows these rectangles constructed on one of these intervals $[x_{i-1}, x_i]$). We shall obtain two stepped figures one of which is contained in a curvilinear trapezoid and the other contains that trapezoid. Upon rotation of the curvilinear trapezoid and the stepped figures, we obtain a body E and two stepped bodies one of which is contained in E and the other contains E. The volumes V_i and V_d of these stepped bodies are equal,

respectively, to

$$\pi \sum_{i=1}^{n} m_i^2 \Delta x_i \text{ and } \pi \sum_{i=1}^{n} M_i^2 \Delta x_i.$$

These expressions are, evidently, the upper and the lower sums of the function $\pi f^2(x)$. Since this function is integrable, the difference between the indicated sums for some partition T of the interval

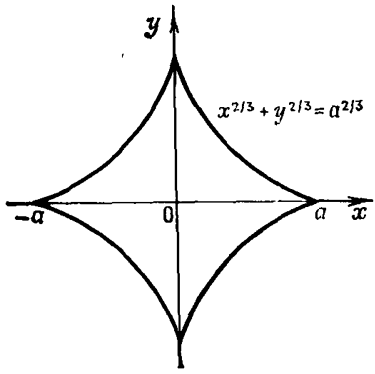


Fig. 11.13

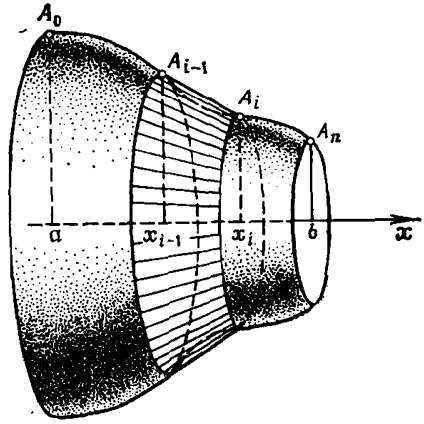


Fig. 11.14

$[a, b]$ will be smaller than the given positive number ε. Consequently, the body E is cubable. Since the limit of the indicated sums is equal to $\pi \int_a^b f^2(x)\,dx$, the volume V of the body E can be found from formula (11.31).

11.3.3. Examples of computing volumes.

1°. The volume of a body obtained by rotating an astroid $x^{2/3} + y^{2/3} = a^{2/3}$ about the Ox axis (Fig. 11.13). Since $y = (a^{2/3} - x^{2/3})^{3/2}$, we have

$$V = \pi \int_{-a}^{a} (x^{2/3} - x^{2/3})^3\,dx = \frac{32}{105}\pi a^3.$$

2°. The volume of a body obtained by rotating a sinusoid on the closed interval $[0, \pi]$ about the Ox axis. We have

$$V = \pi \int_0^{\pi} \sin^2 x\,dx = \pi \int_0^{\pi} \frac{1-\cos 2x}{2}\,dx = \frac{\pi^2}{2}.$$

11.3.4. Area of a surface of revolution. Let us consider a surface Π generated by revolving the graph of the function $y = f(x)$ specified on the closed interval $[a, b]$ about the x-axis (Fig. 11.14). Let

us define the notion of the *squareability* of the surface of revolution Π. Assume that T is the partition of the interval $[a, b]$ by points $a = x_0 < x_1 < \ldots < x_n = b$, and $A_0, A_1, \ldots, A_n$ are the corresponding points of the graph of the function $f(x)$. Let us construct a polygonal line $A_0A_1 \ldots A_n$. Upon rotation of this polygonal line about the axis, we shall obtain a surface $\Pi(A_i)$ formed by the lateral surfaces of truncated cones. Let us designate by $P(x_i)$ the area of the surface $\Pi(A_i)$. If y_i are the ordinates of $f(x)$ at points x_i and l_i is the length of the element $A_{i-1}A_i$ of the polygonal line $A_0A_1 \ldots A_n$, then

$$P(x_i) = 2\pi \sum_{i=1}^{n} \frac{y_{i-1}+y_i}{2} l_i = \pi \sum_{i=1}^{n} (y_{i-1}+y_i)\, l_i. \tag{11.32}$$

Let us formulate the following definition.

1°. *The number P is called the limit of the areas $P(x_i)$ if for any given positive number ε there is a positive number δ such that for any partition T of the closed interval $[a, b]$ the maximum length Δ of whose subintervals is smaller than δ, the inequality $|P(x_i) - P| < \varepsilon$ holds true.*

2°. *The surface of revolution Π is said to be squareable if there exists a limit P of the areas $P(x_i)$. In this case the number P is called the area of the surface Π.*

Let us prove the following assertion.

If the function $f(x)$ possesses a continuous derivative $f'(x)$ on the closed interval $[a, b]$, then the surface Π generated by the rotation of the graph of that function about the Ox axis is squareable and its area P can be computed by the formula

$$P = 2\pi \int_a^b f(x) \sqrt{1+f'^2(x)}\, dx. \tag{11.33}$$

Proof. The length l_i of the element $A_{i-1}A_i$ of the polygonal line $A_0A_1 \ldots A_n$ is equal to $\sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}$. By Lagrange's formula we have $y_i - y_{i-1} = f(x_i) - f(x_{i-1}) = f'(\xi_i)(x_i - x_{i-1})$. Putting $x_i - x_{i-1} = \Delta x_i$, we obtain $l_i = \sqrt{1+f'^2(\xi_i)}\, \Delta x_i$. Therefore, in accordance with (11.32), we have

$$P(x_i) = 2\pi \sum_{i=1}^{n} f(\xi_i) \sqrt{1+f'^2(\xi_i)}\, \Delta x_i + \pi \Big\{ \sum_{i=1}^{n} [(y_{i-1} - f(\xi_i)) + (y_i - f(\xi_i))] \sqrt{1+f'^2(\xi_i)}\, \Delta x_i \Big\}. \tag{11.34}$$

The first sum on the right-hand side of relation (11.34) is the integral sum of the function $2\pi f(x)\sqrt{1+f'^2(x)}$ which, by virtue of the condition stated in the assertion, is integrable and possesses a limit

$P = 2\pi \int_a^b f(x) \sqrt{1 + f'^2(x)}\, dx$. Let us prove that the expression in braces on the right-hand side of relation (11.34) possesses a limit zero. Indeed, suppose ε is any positive number. The function $f(x)$ being uniformly continuous on the closed interval $[a, b]$, we can indicate $\delta > 0$, proceeding from the given ε, such that for $\Delta < \delta$ ($\Delta = \max \Delta x_i$) the inequalities $|y_{i-1} - f(\xi_i)| < \varepsilon$ and $|y_i - f(\xi_i)| < \varepsilon$ hold true. If M is the maximum value of the function $\sqrt{1 + f'^2(x)}$ on the closed interval $[a, b]$ then we get the following estimation for the expression in braces on the right-hand side of relation (11.34):

$$\left|\left\{\sum_{i=1}^{n} [(y_{i-1} - f(\xi_i)) + (y_i - f(\xi_i))] \sqrt{1 + f'^2(\xi_i)}\, \Delta x_i\right\}\right| <$$

$$< 2M\varepsilon \sum_{i=1}^{n} \Delta x_i = 2M(b-a)\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, the limit of the indicated expression is equal to zero. Thus we have proved the existence of the limit P of the areas $P(x_i)$ and established the fact that this limit can be computed by formula (11.33). We have proved the assertion.

Remark 1. The squareability of a surface of revolution can be proved under weaker conditions. It suffices to require that the function $f'(x)$ be defined and integrable on the closed interval $[a, b]$. From this assumption it follows that the function $f(x) \sqrt{1 + f'^2(x)}$ is integrable (see Supplement 1 to Chap. 10). Further arguments do not differ anyhow from the arguments presented in the proof of the assertion given in this subsection.

Remark 2. If the surface π is obtained as a result of rotation of curve L specified by the parametric equations $x = \varphi(t)$, $y = \psi(t)$, $\alpha \leqslant t \leqslant \beta$ about the x-axis, then, making a change of variables under the definite integral sign in formula (11.33), we shall obtain the following expression for the area P of that surface:

$$P = 2\pi \int_\alpha^\beta \psi(t) \sqrt{\varphi'^2(t) + \psi'^2(t)}\, dt. \tag{11.35}$$

Let us consider some examples of computing areas of surfaces of revolution.

1°. Let us find the area P of the surface of the ellipsoid of revolution. Assume that the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ rotates about the Ox axis. We shall first consider the case $a > b$ (rotation about the greater axis of the ellipse). Since in this case $f(x) = \frac{b}{a}\sqrt{a^2 - x^2}$, we shall find, putting $e = \sqrt{\frac{a^2 - b^2}{a^2}}$,

$$P = 2\pi \int_{-a}^{a} f(x) \sqrt{1 + f'^2(x)}\, dx = 2\pi \frac{b}{a} \int_{-a}^{a} \sqrt{a^2 - e^2 x^2}\, dx =$$

$$= 2\pi b \left(b + \frac{a}{e} \arcsin e\right).$$

If $a < b$, then, putting $c = \sqrt{\frac{b^2 - a^2}{b^2}}$ and carrying out the required calculations, we obtain

$$P = 2\pi b \left(b + \frac{a^2}{2b} \frac{1}{e} \ln \frac{1+e}{1-e} \right).$$

2°. Let us find the area P of the surface generated by the rotation about the Ox axis of a cycloid specified by the parametric equations $x = a(t - \sin t)$, $y = a(1 - \cos t)$, $0 \leqslant t \leqslant 2\pi$. From formula (11.35) we have

$$P = 2\pi \int_{\alpha}^{\beta} \psi(t) \sqrt{\varphi'^2(t) + \psi'^2(t)}\, dt = 2\sqrt{2}\, \pi a^2 \int_0^{2\pi} (1 - \cos t)^{3/2}\, dt = \frac{64}{3} \pi a^2.$$

11.4. SOME PHYSICAL APPLICATIONS OF THE DEFINITE INTEGRAL

11.4.1. The mass and the centre of gravity of a nonhomogeneous bar. Let us consider a nonhomogeneous bar located on the closed interval $[a, b]$ of the Ox axis. Suppose $\rho(x)$ is the linear density of the bar*. We shall designate as T the partition of the interval $[a, b]$ by points $a = x_0 < x_1 < \ldots < x_n = b$. Let us choose a point ξ_i on each subinterval $[x_{i-1}, x_i]$ and compile a sum $\sum_{i=1}^{n} \rho(\xi_i)\, \Delta x_i$. Since every addend of this sum is an approximate value of the mass of a part of the bar on the subinterval $[x_{i-1}, x_i]$, it is natural to take the indicated sum as an approximate value of the mass of the whole bar. In accordance with this preliminary argument we shall *determine the mass M of the whole bar as the limit of the sums* $\sum_{i=1}^{n} \rho(\xi_i)\, \Delta x_i$ *when* $\Delta = \max \Delta x_i$ *tends to zero, that is, as the integral* $\int_a^b \rho(x)\, dx$. Thus we have

$$M = \int_a^b \rho(x)\, dx. \tag{11.36}$$

To determine the centre of gravity of a nonhomogeneous bar, we shall make use of the formula for the coordinate of the centre of gravity of the system $\{m_i(x_i)\}$ of particles having the masses m_i and located at points x_i of the x-axis. Namely, the coordinate x_c of the centre of gravity of the system $\{m_i\}$ can be found by the

* If Δm is the mass of a part of the bar on the subinterval $[x, x + \Delta x]$ then the ratio $\frac{\Delta m}{\Delta x}$ is called the mean linear density of the bar on that subinterval. The linear density $\rho(x)$ is the limit $\rho(x) = \lim\limits_{\Delta x \to 0} \frac{\Delta m}{\Delta x}$.

formula

$$x_c = \frac{m_1x_1 + m_2x_2 + \ldots + m_nx_n}{m_1 + m_2 + \ldots + m_n} = \frac{\sum\limits_{i=1}^{n} m_ix_i}{\sum\limits_{i=1}^{n} m_i}. \tag{11.37}$$

Let us discuss partition T of the closed interval $[a, b]$ by points $a = x_0 < x_1 < \ldots < x_n = b$ and calculate mass m_i of a part of the bar located on the subinterval $[x_{i-1}, x_i]$. From formula (11.36) we have $m_i = \int\limits_{x_{i-1}}^{x_i} \rho(x)\,dx$. Applying mean-value formula (10.13), we see also that $m_i = \rho(\xi_i)\,\Delta x_i$. Assuming that the mass m_i is concentrated at point ξ_i of the subinterval $[x_{i-1}, x_i]$, we can consider the nonhomogeneous bar as a system of particles with masses m_i located at points ξ_i of the interval $[a, b]$. Since

$$\sum_{i=1}^{n} m_i = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} \rho(x)\,dx = \int_a^b \rho(x)dx = M,$$

we shall find from formula (11.37) an approximate expression for the coordinate x_c of the centre of gravity of the nonhomogeneous bar

$$x_c \approx \frac{\sum\limits_{i=1}^{n} \xi_i\rho(\xi_i)\,\Delta x_i}{M}. \tag{11.38}$$

The expression in the numerator of the right-hand side of (11.38) is the integral sum for the function $x\rho(x)$ on the interval $[a, b]$. In accordance with our reasoning, we shall determine the coordinate x_c of the centre of gravity of the nonhomogeneous bar by the formula

$$x_c = \frac{\int\limits_a^b x\rho(x)\,dx}{\int\limits_a^b \rho(x)\,dx}. \tag{11.39}$$

11.4.2. The work performed by a variable force. Suppose a particle moves from point a of the x-axis to point b of that axis under the force F parallel to the Ox axis. We shall assume that this force is a function of x defined on the closed interval $[a, b]$. Suppose T is the partition of the interval $[a, b]$ by points $a = x_0 < x_1 < \ldots < x_n = b$. Let us choose a point ξ_i on each subinterval $[x_{i-1}, x_i]$ and assume the expression $\sum\limits_{i=1}^{n} F(\xi_i)\,\Delta x_i$ to be an approximate value of the work A of the variable force $F(x)$ on the interval $[a, b]$.

In good agreement with our previous reasoning, we shall define the work A of the variable force $F(x)$ on the interval $[a, b]$ as the integral $\int_a^b F(x)\,dx$. Thus we have

$$A = \int_a^b F(x)\,dx. \tag{11.40}$$

SUPPLEMENT TO CHAPTER 11

AN EXAMPLE OF A NONSQUAREABLE FIGURE

1. We shall use the term *half-open triangle* for the set of points of a triangle whose boundary* lacks the points of its two sides and two vertices adjoining these sides. Let us consider the construction of a curve L which will be a part of the

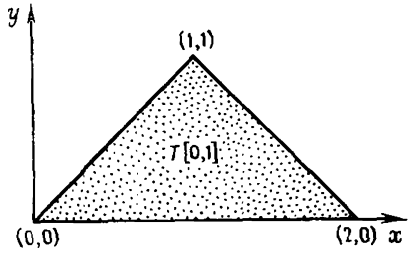


Fig. 11.15

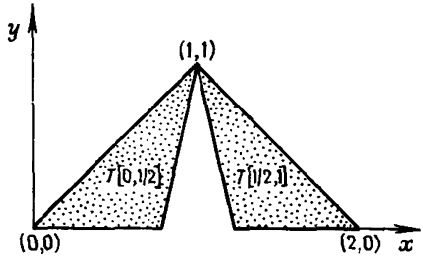


Fig. 11.16

boundary of a nonsquareable figure Q. This construction is carried out by way of successive deletions of certain half-open triangles from some given isosceles right-angled triangle T which we shall denote by T [0, 1] to make our further reasoning more convenient. The coordinates of the vertices of the triangle are equal to (0, 0), (1, 1), (2, 0) (Fig. 11.15). We shall describe now the process of successive deletions of certain half-open triangles from the triangle T [0, 1].

1. A half-open triangle is deleted whose one vertex has the coordinates (1, 1) and two others are located on the x-axis. The area S_1 of the triangle being removed is equal to $\frac{1}{4}$. The resulting figure is shown in Fig. 11.16. It consists of two triangles $T\left[0, \frac{1}{2}\right]$ and $T\left[\frac{1}{2}, 1\right]$ whose areas are equal to each other.

2. One triangle is deleted from each of the triangles $T\left[0, \frac{1}{2}\right]$ and $T\left[\frac{1}{2}, 1\right]$, the sum S_2 of the areas of the removed triangles being equal to $\frac{1}{8}$. The resulting figure is illustrated in Fig. 11.17. It consists of four triangles $T\left[0, \frac{1}{4}\right]$, $T\left[\frac{1}{4}, \frac{1}{2}\right]$, $T\left[\frac{1}{2}, \frac{3}{4}\right]$ and $T\left[\frac{3}{4}, 1\right]$ whose areas are equal to each other.

* The boundary of a triangle is the set of points of its sides and vertices.

3. One triangle is removed from each of the indicated triangles, the sum S_3 of the removed triangles being equal to $\frac{1}{16}$. The resulting figure is shown in

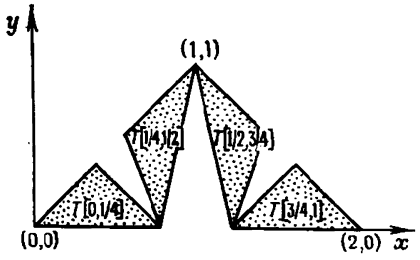


Fig. 11.17

Fig. 11.18. It consists of eight triangles:

$$T\left[0, \frac{1}{8}\right], \ T\left[\frac{1}{8}, \frac{1}{4}\right], \ T\left[\frac{1}{4}, \frac{3}{8}\right], \ T\left[\frac{3}{8}, \frac{1}{2}\right],$$
$$T\left[\frac{1}{2}, \frac{5}{8}\right], \ T\left[\frac{5}{8}, \frac{3}{4}\right], \ T\left[\frac{3}{4}, \frac{7}{8}\right], \ T\left[\frac{7}{8}, 1\right],$$

whose areas are equal to each other.

4. One triangle is removed from each of the indicated triangles, the sum S_4 of the removed triangles being equal to $\frac{1}{32}$. The resulting figure is illustrated

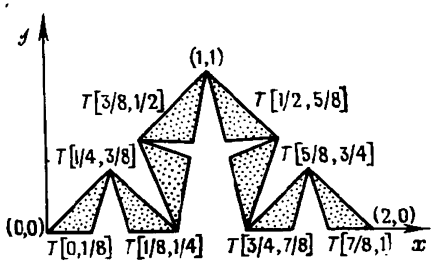


Fig. 11.18

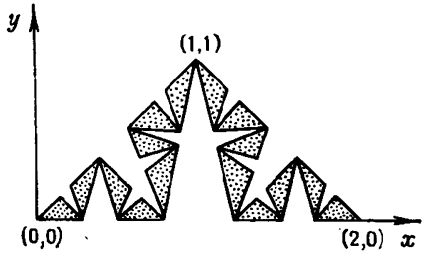


Fig. 11.19

in Fig. 11.19. It consists of sixteen triangles of equal areas. We shall designate each of these triangles by the symbol

$$\left\{T\left[\frac{p}{2^4}, \frac{p+1}{2^4}\right], \quad p=0, 1, \ldots, 15.\right.$$

The subsequent procedure of removing triangles is obvious. Let us now specify the curve L. The triangles $T\left[\frac{p}{2^n}, \frac{p+1}{2^n}\right]$ (p and n are any non-negative integers satisfying the condition $p < 2^n$) obtained in the process described above possess the following property: suppose $T\left[\frac{p_1}{2^{n_1}}, \frac{p_1+1}{2^{n_1}}\right]$

and $T\left[\frac{p_2}{2^{n_2}}, \frac{p_2+1}{2^{n_2}}\right]$ are two triangles such that $\frac{p_1}{2^{n_1}} \leqslant \frac{p_2}{2^{n_2}} < \frac{p_2+1}{2^{n_2}} \leqslant \frac{p_1+1}{2^{n_1}}$. Then the second of the triangles is contained in the first. Note the following obvious property of the triangles $T\left[\frac{p}{2^n}, \frac{p+1}{2^n}\right]$: their diameters * tend to zero as $n \to \infty$. Assume that $\left\{T\left[\frac{p_k}{2^{n_k}}, \frac{p_k+1}{2^{n_k}}\right]\right\}$, $k = 1, 2, \ldots$, is a *nested system of triangles* (this means that the triangle corresponding to the subscript k contains the triangle corresponding to the subscript $k+1$, and that at $k \to \infty$ the diameters of the triangles tend to zero). Each *nested system of triangles of this kind has strictly one point in common***. Let us consider various nested systems of the triangles we have been discussing. *The curve L will be defined as the set $\{M\}$ of various points each of which is a common point of some nested system of the triangles $T\left[\frac{p}{2^n}, \frac{p+1}{2^n}\right]$ indicated above.*

It should be pointed out that the vertices of all the triangles $T\left[\frac{p}{2^n}, \frac{p+1}{2^n}\right]$ belong to the set $\{M\}$ (curve L). To verify this fact, take into consideration that the vertex of each of these triangles belongs to the nested system of the triangles $\left\{T\left[\frac{2^k p}{2^{n+k}}, \frac{2^k p+1}{2^{n+k}}\right]\right\}$ and the system $\left\{T\left[\frac{2^k p-1}{2^{n+k}}, \frac{2^k p}{2^{n+k}}\right]\right\}$. To ascertain the fact that the set $\{M\}$ we have constructed is a simple curve in the sense of the definition given in 1.1 of this chapter, we must prove that all the points of the set $\{M\}$ are specified by the parametric equations $x = \varphi(t)$, $y = \psi(t)$, $a \leqslant t \leqslant b$, where $\varphi(t)$ and $\psi(t)$ are continuous cunctions ***.

Let us consider a closed interval [0, 1] of the t-axis. We shall associate each interval $\left[\frac{p}{2^n}, \frac{p+1}{2^n}\right]$, where p and n are any nonnegative integers, $p < 2^n$, with a triangle $T\left[\frac{p}{2^n}, \frac{p+1}{2^n}\right]$****. Figure 11.20 shows the closed intervals with which the triangles $T\left[\frac{p}{2^3}, \frac{p+1}{2^3}\right]$ are associated. Any point t of the closed interval [0, 1] belongs to all the closed intervals of

---

* The term diameter of a triangle is used for the length of its maximum side.

** We have proved in Chapter 3 (see 3.3.2) that the nested system of closed intervals has strictly one point in common. Mapping a nested system of triangles on the Ox and Oy axes, we shall obtain a nested system of closed intervals on the coordinate axes. Assume that x and y are the common points of the indicated nested systems of closed intervals on the Ox and Oy axes respectively. The reader can easily verify that point M with coordinates x and y is the only common point of the nested system of triangles being considered.

*** The fact that various t are associated with various points of the set $\{M\}$ is evident from the construction of the curve L.

**** Note that each triangle of this kind is associated with only one closed interval $\left[\frac{p}{2^n}, \frac{p+1}{2^n}\right]$.

some nested system $\left\{\left[\frac{p_k}{2^{n_k}}, \frac{p_k+1}{2^{n_k}}\right]\right\}$ of closed intervals *. Let us associate with this point t a common point M of the nested system of triangles $\left\{T\left[\frac{p_k}{2^{n_k}}, \frac{p_k+1}{2^{n_k}}\right]\right\}$. Thus we associate each value of t from the closed interval [0, 1] with two numbers x and y which are the coordinates of point M. Consequently, x and y are the functions of parameter t. Let

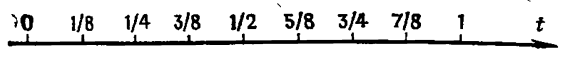


Fig. 11.20

us verify that these functions $x=\varphi(t)$ and $y=\psi(t)$ are continuous on the closed interval [0, 1]. Indeed, assume that ε is any given positive number, t is a given point belonging to the closed interval [0, 1] and M is a point belonging to the curve L and specified by this value of parameter t. From the nested system $\left\{T\left[\frac{p_k}{2^{n_k}}, \frac{p_k+1}{2^{n_k}}\right]\right\}$ of triangles, specifying point M, we shall select a triangle whose diameter is smaller than ε and consider a closed interval $\left[\frac{p_k}{2^{n_k}}, \frac{p_k+1}{2^{n_k}}\right]$ which contains point t specifying point M (and, hence, x and y). All the points of the curve L specified by the values of t from this closed interval are located in the indicated triangle and, therefore, their coordinates differ from those of point M by not more than ε. But that precisely means that the functions $\varphi(t)$ and $\psi(t)$ are continuous at the point we have indicated.

2. Let us now pass to constructing a nonsquareable figure Q. We shall consider a square $\underline{Q}$ whose side is equal to 2. On each side of this square we shall construct isosceles right-angled triangles T_1, T_2, T_3, T_4 and, as a result, we shall obtain a square $\overline{Q}$ with side equal to $2\sqrt{2}$ (Fig. 11.21). Then from each of these triangles we shall delete half-open triangles as was described above in Section 1. We shall obtain a figure Q bounded by a closed curve consisting of four curves congruent** to the curve L (see Sec. 1). Let us prove that the figure Q we have obtained is nonsquareable. We shall consider two particular sequences of polygons $\{\underline{Q_n}\}$ and $\{\overline{Q_n}\}$, the first of which consists of polygons inscribed into the

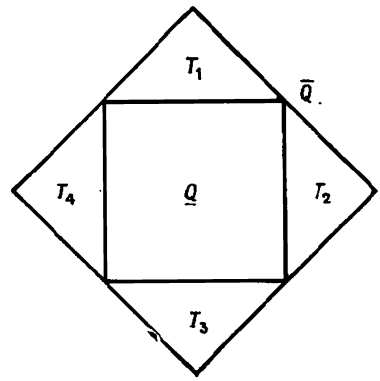


Fig. 11.21

* Suppose t is any point of the closed interval [0, 1] and n is any positive integer. Then, evidently, point t belongs to a certain closed interval $\left[\frac{p}{2^n}, \frac{p+1}{2^n}\right]$, with each interval of that kind, corresponding to the number $n+1$, being contained in the interval corresponding to the number n.

** The sets A and B are said to be congruent if they can be superimposed by being shifted.

figure Q and the second consists of polygons circumscribed about Q. The sequence $\{\overline{Q}_n\}$ is produced by adding to the square $\overline{Q}$ the half-open triangles removed from the triangles T_1, T_2, T_3, T_4 at each odd stage of the process described in 11.1. The sequence $\{\underline{Q}_n\}$ is obtained by removing from the square $\overline{Q}$ half-open triangles deleted from the triangles T_1, T_2, T_3, T_4 at each even stage of the process described in Sec. 1. It is evident that any polygon inscribed into the figure Q is contained in some polygon $\underline{Q}_n$ and any polygon circumscribed about Q contains some polygon $\overline{Q}_n$. Therefore, the limit of the sequence $\{\underline{S}_n\}$ of the areas of the polygons $\underline{Q}_n$ is equal to the *lower area* $\underline{P}$ of the figure Q and the limit of the sequence $\{\overline{S}_n\}$ of the areas of the polygons $\overline{Q}_n$ is equal to the *upper area* $\overline{P}$ of the figure Q. It is easy to verify that $\underline{S}_n = 4 + \sum_{k=1}^{n} \frac{1}{4^{k-1}}$, and $\overline{S}_n = 8 - \frac{1}{2}\sum_{k=1}^{n} \frac{1}{4^{k-1}}$*. Therefore, $\underline{P} = \lim_{n\to\infty} \underline{S}_n = \frac{16}{3}$ and $\overline{P} = \lim_{n\to\infty} \overline{S}_n = \frac{22}{3}$. Since $\overline{P} \neq \underline{P}$, the *figure Q is nonsquareable.* Note that the difference $\overline{P} - \underline{P} = 2$. Thus the boundary of the figure Q under consideration has the area equal to 2.

3. Let us prove that *any part of the curve L bounded by two distinct points is not rectifiable.* We shall first prove that such a part L' of the curve L has an area different from zero, that, is, any polygon covering L' has an area exceeding some positive number. Note that L' contains part L'' corresponding to the points of some closed interval $\left[\frac{p}{2^n}, \frac{p+1}{2^n}\right]$ and, therefore, L'' is contained in the triangle $T\left[\frac{p}{2^n}, \frac{p+1}{2^n}\right]$ and can be obtained by deleting from this triangle some half-open triangles (see Sec. 1 of the present supplement). It is easy to calculate that the sum S of the areas of all half-open triangles being removed is smaller than the area S_T of the triangle $T\left[\frac{p}{2^n}, \frac{p+1}{2^n}\right]$. Consequently, the part L'' has an area equal to $S_T - S > 0$. When proving, in 11.2 the squareability of a figure bounded by a rectifiable curve, we have proved that the area of a rectifiable curve is equal to zero (a rectifiable curve can be covered by a polygonal line of an arbitrarily small area). Therefore, part L'' of the curve L and, consequently, part L' containing L'' are nonrectifiable either.

Remark. Each of the functions $\varphi(t)$ and $\psi(t)$ we have constructed *does not have a derivative at any point of the closed interval* $[0, 1]$.

---

* We can easily obtain these formulas if we take into account that the sums of the areas of the triangles removed at odd stages of the process produce a geometric progression $1, \frac{1}{4}, \ldots$, and the sums of the area of the triangles removed at even stages of the process produce a geometric progression $\frac{1}{2}, \frac{1}{8}, \ldots$.

CHAPTER 12

APPROXIMATE METHODS OF CALCULATING ROOTS OF EQUATIONS AND DEFINITE INTEGRALS

In this chapter we deal with approximate methods of calculating roots of algebraic and transcendental equations and computing definite integrals.

12.1. APPROXIMATE METHODS OF CALCULATING ROOTS OF EQUATIONS

We shall be occupied here with approximate calculations of one of the roots of the equation $f(x) = 0$, where $y = f(x)$ is some continuous or differentiable function. We shall take it as a fact that the root c we are interested in is isolated on a certain closed interval $[a, b]$, that is, we shall assume that this root is an interior point of the interval $[a, b]$ which does not contain any other roots of the equation being discussed. In practical applications the size of the interval $[a, b]$ is usually determined by a rough estimation*.

12.1.1. The "fork" method. We shall begin with a method frequently used for approximate calculations of roots on modern computers.

Suppose that the root c of the equation $f(x) = 0$ we are interested in is isolated on some closed interval $[a, b]$. As concerns the function $f(x)$, we shall assume that it is continuous on the interval $[a, b]$ and possesses values of unlike signs at the end-points of the interval. In what follows, for the sake of brevity, we shall term as a *"fork"* every closed interval at whose end-points $f(x)$ possesses values of unlike signs.

We shall now describe the method of finding the roots of the equation $f(x) = 0$ called a *"fork" method.*

For definiteness we shall assume that $f(a) < 0$, $f(b) > 0$. If we divide the interval $[a, b]$ in half, two cases may occur: (1) the value of the function at the middle point of the interval $[a, b]$ is equal to zero (in that case the required root is obtained), (2) the indicated value is nonzero. In this case one half of the closed interval is a "fork". We shall designate this half as $[a_1, b_1]$. It is evident that $f(a_1) < 0$, $f(b_1) > 0$. We shall deal with the interval $[a_1, b_1]$ in the same way as with the interval $[a, b]$, i.e. divide it in half. Proceeding with analogous reasoning, we shall have two possibil-

* In this case use can be made of some additional information on the position of the root that can be derived from the physical content of the problem.

ities: (1) either the process described above will terminate because in the middle of some interval the value of the function will turn out to be zero (in that case the required root is obtained), (2) or the process described can be continued indefinitely and we shall receive a nested system of closed intervals—forks—$[a_1, b_1]$, $[a_2, b_2]$,... ..., $[a_n, b_n]$, ..., with $f(a_n) < 0$, $f(b_n) > 0$ for any number n. This nested system of closed intervals has one common point c to which each of the sequences $\{a_n\}$ and $\{b_n\}$ converges (see the corollary of Theorem 3.15). Let us prove that c is the required root, i.e. $f(c) = 0$. Since the function $f(x)$ is continuous at point c, each of the sequences $\{f(a_n)\}$ and $\{f(b_n)\}$ converges to $f(c)$. But then we obtain from the conditions $f(a_n) < 0$, $f(b_n) > 0$, by virtue of Theorem 3.13 and the remark to that theorem, that two inequalities $f(c) \leqslant 0$ and $f(c) \geqslant 0$ are simultaneously valid, i.e. $f(c) = 0$.

The above reasonings give an algorithm for searching the required root c. We can take the point $\frac{a_n+b_n}{2}$, that is the middle point of the interval $[a_n, b_n]$, as an approximate value of the root. The length of the interval $[a_n, b_n]$ being equal to $\frac{b-a}{2^n}$, the number $\frac{a_n+b_n}{2}$ differs from the exact value of the root by not more than $\frac{b-a}{2^{n+1}}$. Thus the process of a successive division of fork-intervals in half makes it possible to compute the sought root c with any preassigned accuracy. Since this process requires multiple repetitions of *the same kind* of computing operations, it is most convenient to carry out the operations on high-speed mathematical computers.

12.1.2. The tangent method*. This is one of the most efficient methods of approximate calculation of the roots of the equation $f(x) = 0$.

Assume that the desired root c of the equation $f(x) = 0$ is isolated on the closed interval $[a, b]$. We shall describe the method without elucidating, for the time being, the conditions under which this method is applicable.

Let us consider the graph of the function $f(x)$ on the closed interval $[a, b]$ (Fig. 12.1). We shall take a certain value x_0 from the interval $[a, b]$ for the zero approximation of the required root and denote by B_0 the point of the graph of the function with the abscissa x_0. Let us draw through point B_0 a tangent to the graph of the function and take the abscissa x_1 of the point of intersection of the tangent with the x-axis** as a first approximation of the desired

* It is also known as *Newton's method*.

** Since the tangent at point B_0 is the graph of the differential of the function $y = f(x)$ at point x_0, the indicated method of finding a first approximation x_1 is based on *replacing the function by its differential* at point x_0.

root (see Fig. 12.1). Next we shall draw a tangent to the graph of the function through point B_1 with the abscissa x_1 and take the abscissa x_2 of the point of intersection of the tangent line with the x-axis as a second approximation. Continuing this process indefinitely, we shall construct a sequence $x_0, x_1, \ldots, x_n, \ldots$ of the approximate values of the desired root.

In practical applications it is more convenient to obtain a recurrence formula expressing x_{n+1} through x_n. For that purpose it is

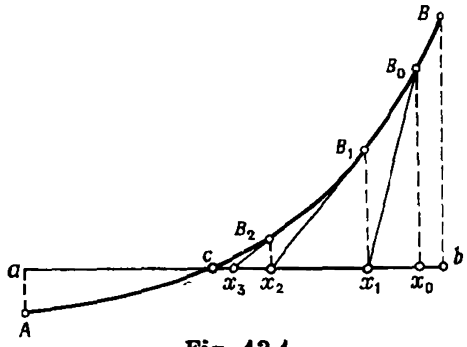


Fig. 12.1

necessary to take the equation $Y - f(x_n) = f'(x_n)(x - x_n)$ of the tangent line to the graph of the function at point B_n and calculate the abscissa x_{n+1} of the point of intersection of that tangent with the x-axis. We obtain

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{12.1}$$

Formula (12.1) determines the algorithm of the tangent method. Thus, the tangent method is the method of successive approximations (iteration method) constructed by means of the recurrence formula (12.1). Our further aim is to substantiate the tangent method.

We shall elucidate in 12.1.5 the conditions under which the sequence of values x_n determined by formula (12.1) converges to the desired root c and shall estimate the error, that is, the deviation of the approximate value x_n from the exact value of the root c.

12.1.3. The chord method. The method of chords is one of the most favourable methods of approximation used for solving the equation $f(x) = 0$.

We shall now describe the method without establishing the conditions under which it is applicable.

Assume that the required root c of the equation $f(x) = 0$ is isolated on the closed interval $[a, b]$ and consider the graph of the function $f(x)$ on that interval (Fig. 12.2). Let us take some number x_0 from the interval $[a, b]$ as the zero approximation of the desired root and designate the points of the graph of the function with the

abscissas x_0 and b as A_0 and B. Let us draw a chord A_0B through points A_0 and B of the graph of the function and take the abscissa x_1 of the point of intersection of that chord with the x-axis as a first approximation of the desired root (see Fig. 12.2). Then we shall draw a chord through the points A_1, with the abscissa x_1, and B of the graph of the function. The abscissa x_2 of the point of intersection of the chord A_1B with the x-axis will be taken as a second ap-

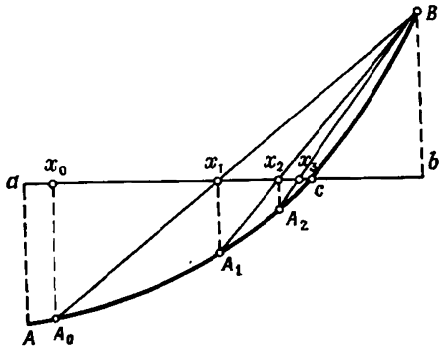


Fig. 12.2

proximation. Continuing this process indefinitely, we shall construct a sequence $x_0, x_1, x_2, \ldots, x_n, \ldots$ of the approximate values of the desired root.

In practical applications it is more convenient to obtain a recurrence formula expressing x_{n+1} in terms of x_n. For that purpose we shall take the equation $\frac{Y-f(x_n)}{f(b)-f(x_n)}=\frac{x-x_n}{b-x_n}$ of the chord passing through points $A_n(x_n, f(x_n))$ and $B(b, f(b))$, and calculate the abscissa x_{n+1} of the point of intersection of that chord with the x-axis. We shall obtain

$$x_{n+1}=x_n-\frac{(b-x_n)f(x_n)}{f(b)-f(x_n)}. \tag{12.2}$$

Formula (12.2) determines the algorithm of the chord method. Thus, the chord method is the method of iterations which can be constructed by means of the recurrence formula (12.2). Our further aim is to substantiate the method of chords.

In 12.1.6 we shall elucidate the conditions under which the sequence of the values x_n converges to the desired root c and estimate the error of the chord method.

12.1.4. The method of iterations (successive approximations). It is clear from 12.1.2 and 12.1.3 that the tangent method and the chord method are connected by the common idea of the construction of successive approximations of the required root. This very

idea serves as the basis for the method described in the present subsection.

This method will be considered as applied to the equation

$$x = F(x). \tag{12.3}$$

We shall now introduce the concept of the *iterative sequence.*

We shall say that the sequence $x_0, x_1, \ldots, x_n, \ldots$ is *iterative* if for any $n \geqslant 1$ the term x_n can be expressed through the term x_{n-1} by means of the recurrence formula $x_n = F(x_{n-1})$, and any number from the domain of definition of the function $F(x)$ is taken as x_0. We shall prove that under certain conditions the iterative sequence converges to the root of equation (12.3) and, consequently, its terms can be taken as approximate values of that root.

The following assertion is valid.

Assertion 1. *Suppose that the function $F(x)$ is continuous on the closed interval $[a, b]$ and assume that all the terms o the iterative sequence $x_0, x_1, \ldots, x_n, \ldots$ belong to that interval. Then, if that sequence converges to a certain number c, the indicated number c is a root of equation* (12.3).

Proof. Since the sequence $\{x_n\}$ converges to c and all its terms belong to $[a, b]$, the limit c also belongs to the interval $[a, b]$ (see Corollary 2 of Theorem 3.13). By the hypothesis, the function $F(x)$ is continuous at point c, and, therefore, the sequence $\{F(x_{n-1})\}$ converges to $F(c)$. Thus, in the limit, as $n \to \infty$, the equality $x_n = F(x_{n-1})$ passes into the equality $c = F(c)$, that is, c is a root of equation (12.3). The assertion we have proved will be widely applied in 12.1.5. and 12.1.6 to substantiate the methods of tangents and chords.

We shall now prove one more assertion frequently used for approximate calculations of the root of equation (12.3) with the aid of iterative sequence.

Assertion 2. *Let c be a root of equation* (12.3) *and assume that the derivative of the function $F(x)$ complies with the condition $|F'(x)| \leqslant \alpha < 1$ on some closed interval $[c - \varepsilon, c + \varepsilon]$ symmetric with respect to point c. Then, the iterative sequence $x_0, x_1, \ldots, x_n, \ldots$, in which any number from the interval $[c - \varepsilon, c + \varepsilon]$ is taken as x_0, converges to the indicated root c.*

Proof. We shall prove first of all that all the terms of the iterative sequence $\{x_n\}$ belong to the indicated interval $[c - \varepsilon, c + \varepsilon]$. Actually, x_0 belongs to that interval by the hypothesis. Therefore, under the assumption that x_{n-1} belongs to that interval, it is sufficient to prove that x_n belongs to it as well. For that purpose we shall apply Lagrange's formula to the difference $F(x_{n-1}) - F(c)$ and take into account that $F(c) = c$, $x_n = F(x_{n-1})$. We shall have

$$x_n - c = F(x_{n-1}) - F(c) = F'(\xi)(x_{n-1} - c), \tag{12.4}$$

where ξ is some point lying between x_{n-1} and c and hence belonging to the closed interval $[c-\varepsilon, c+\varepsilon]$. Since $|F'(\xi)| \leqslant \alpha < 1$, we obtain from equality (12.4)

$$|x_n - c| \leqslant \alpha |x_{n-1} - c|. \tag{12.5}$$

Since $0 < \alpha < 1$, we obtain from (12.5)

$$|x_n - c| < |x_{n-1} - c|. \tag{12.6}$$

Inequality (12.6) establishes the fact that each succeeding term x_n is located closer to c than the preceding term x_{n-1} and, hence, since x_{n-1} belongs to the closed interval $[c-\varepsilon, c+\varepsilon]$ and since this interval is symmetric with respect to c, x_n also belongs to that interval. It remains to prove that the sequence $\{x_n\}$ converges to c. Inequality (12.5) being valid for all the numbers n, we can use this inequality to obtain

$$|x_n - c| \leqslant \alpha^n |x_0 - c|. \tag{12.7}$$

It is evident from the last inequality that $x_n \to c$ since $\alpha^n \to 0$. We have proved assertion 2.

Practical remarks are due concerning the assertion we have just proved. Let us assume that by means of rough estimation we have established the fact that

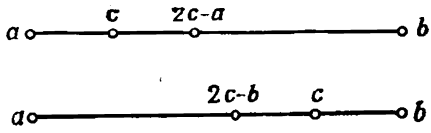


Fig. 12.3

the root of equation (12.3) we are interested in is isolated on some closed interval $[a, b]$ on which the derivative of the function $F(x)$ complies with the condition $|F'(x)| \leqslant \alpha < 1$.

Since in the general case the interval $[a, b]$ *is not symmetric* with respect to the required root, the question naturally arises how to choose the zero approximation x_0 to be able to apply Assertion 2 we have proved. Note that wherever the required root c is within the interval $[a, b]$, at least one of the intervals $[a, 2c-a]$ and $[2c-b, b]$ (Fig. 12.3) symmetric with respect to c *belongs entirely to the closed interval* $[a, b]$. Therefore, at least one of the points a and b belongs to the closed interval, symmetric with respect to the root c, throughout which $|F'(x)| \leqslant \alpha < 1$. Hence, at least one of the points a and b can be taken, in accordance with Assertion 2 proved above, as x_0. More specifically, that of the points a and b should be taken as x_0 for which the approximation $x_1 = F(x_0)$ is confined to the closed interval $[a, b]$.

The case most frequently met in practical applications is the derivative $F'(x)$ possessing a definite sign on the closed interval $[a, b]$. If the sign is positive then it follows from formula (12.4) that the sequence $\{x_n\}$ is monotonic. This case leads to the so-called *stepped diagram* depicted in Fig. 12.4. Now if the derivative $F'(x)$ is nega-

tive on the interval $[a, b]$, then it can be seen from the same formula (12.4) that any two consecutive terms x_{n-1} and x_n lie on different sides of the root c. This case leads to the so-called *spiral diagram* shown in Fig. 12.5.

Remark. The question arises as to the evaluation of the error of the iteration method, that is, the evaluation of the deviation of an

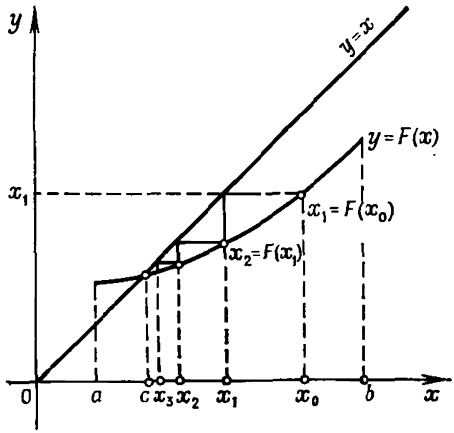


Fig. 12.4

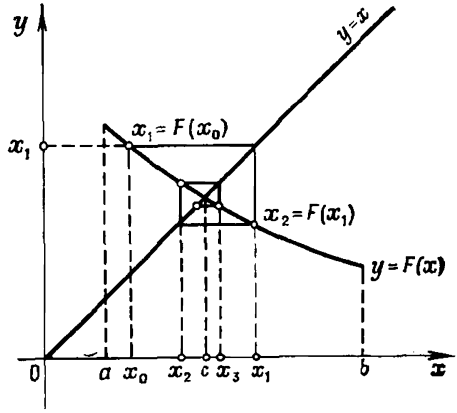


Fig. 12.5

nth approximation x_n from the exact value of the root c. Formula (12.7) immediately yields the following estimation:

$$|x_n - c| \leqslant \alpha^n (b - a),$$

where α is the supremum of the function $|F'(x)|$ on the interval $[a, b]$ on which the root being considered is isolated. If the derivative $F(x)$ is negative on the closed interval $[a, b]$ then, as has been indicated above, x_{n-1} and x_n lie on different sides of the root c and, therefore, the following estimate is valid:

$$|x_n - c| \leqslant |x_n - x_{n-1}|.$$

Now if, in the case being considered, the half-sum of the two successive approximations

$$x_n^* = \frac{x_n + x_{n-1}}{2}$$

is taken as an approximate value of the root, the following estimation of the error ensues:

$$|x_n^* - c| \leqslant \frac{|x_n - x_{n-1}|}{2}.$$

12.1.5. Substantiation of the tangent method.

1°. We shall first consider the case where the desired root c of the equation $f(x) = 0$ is isolated on some closed interval $[a, b]$ on which the function $f(x)$ has *a nonvanishing first derivative and a bounded second derivative*. Let us prove that in this case there can be found a sufficiently small neighbourhood of the root c such that if the zero approximation x_0 is in this neighbourhood, then the sequence $\{x_n\}$, specified by the recurrence formula (12.1), converges to the root c.

We must note first of all that the equation

$$x = F(x), \quad \text{where } F(x) = x - \frac{f(x)}{f'(x)} \tag{12.8}$$

possesses on the interval $[a, b]$ only one root c coinciding with the root of the equation $f(x) = 0$. Therefore, we shall solve equation (12.8) instead of the equation $f(x) = 0$. For that purpose, having taken some x_0, we shall construct an iterative sequence

$$x_{n+1} = F(x_n) = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{12.9}$$

Take note of the fact that the recurrence formula (12.9) exactly coincides with the recurrence formula (12.1).

To ascertain the convergence of the iterative sequence $\{x_n\}$ to the required root c, it is sufficient to prove that in some ε-neighbourhood of the root c the derivative $F'(x)$ satisfies the condition $|F'(x)| \leqslant \alpha < 1$ and to take x_0 in the indicated ε-neighbourhood (see Assertion 2 in 12.1.4). By virtue of the requirement imposed on the function $f(x)$, there can be found positive numbers m and N such that everywhere on the interval $[a, b]$ there hold inequalities

$$|f'(x)| \geqslant m > 0^*, \quad |f''(x)| \leqslant N. \tag{12.10}$$

Since we have

$$F'(x) = 1 - \frac{[f'(x)]^2 - f(x) f''(x)}{[f'(x)]^2} = \frac{f(x) f''(x)}{[f'(x)]^2},$$

the following evaluation follows from inequality (12.10):

$$|F'(x)| \leqslant \frac{|f(x)| N}{m^2}. \tag{12.11}$$

It follows from the continuity of the function $f(x)$ that in some ε-neighbourhood of the root c this function satisfies an inequality

$$|f(x)| \leqslant \frac{m^2}{N} \alpha, \tag{12.12}$$

---

* These inequalities follow from the fact that the derivative $f'(x)$ is continuous and does not vanish on the interval being considered.

where α is a fixed number belonging to the interval $0 < \alpha < 1$. Comparing inequalities (12.11) and (12.12), we see that everywhere in the indicated ε-neighbourhood of the root

$$| F'(x) | \leqslant \alpha < 1.$$

Thus we have proved the convergence of sequence (12.9) to the root c.

Remark 1. We have proved the convergence of the sequence $\{x_n\}$ to the root c only under the condition that the zero approximation x_0 lies within a sufficiently small ε-neighbourhood of the root c. The choice of the needed x_0 can be easily made on a modern computer as a result of several trials.

Remark 2. Let us estimate the deviation of the approximate value x_{n+1} of the root from its exact value c. With that aim in view, let us expand the function $f(x)$ in the neighbourhood of x_n by Taylors's formula with the remainder in Lagrange's form:

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{1}{2} f''(\xi)(x - x_n)^2.$$

Putting $x = c$ in this formula and taking into account that $f(c) = 0$, we have

$$0 = f(x_n) + f'(x_n)(c - x_n) + \frac{1}{2} f''(\xi)(c - x_n)^2.$$

Subtracting from the last formula the equation $f(x_n) + f'(x_n) \times (x_{n+1} - x_n) = 0$ following from recurrence relation (12.9), we obtain

$$x_{n+1} - c = \frac{1}{2} \frac{f''(\xi)}{f'(x_n)} (x_n - c)^2.$$

From this, making use of notation (12.10) accepted above, we arrive at the following inequality:

$$|x_{n+1} - c| \leqslant \frac{N}{2m} |x_n - c|^2.$$

Employing this evaluation consecutively for $n = 0, 1, 2, \ldots$, we obtain the following estimation:

$$|x_{n+1} - c| \leqslant \left(\frac{N}{2m}\right)^n |x_0 - c|^{2n}.$$

2°. We shall now substantiate the tangent method under somewhat different assumptions.

Assume that the desired root c of the equation $f(x) = 0$ is isolated on the closed interval $[a, b]$ on which $f(x)$ has a *monotonic and continuou first derivative retaining a definite sign.*

We shall assume for definiteness that the derivative *is nondecreasing and positive* on the interval $[a, b]$. Let us prove that the iterative sequence $\{x_n\}$, in which $x_0 = b$ and x_{n+1} is defined in terms of x_n with the aid of formula (12.9), converges to the root c. For our pur-

pose it is sufficient to prove that all x_n belong to the interval $[a, b]$ and that the sequence $\{x_n\}$ is convergent (see Assertion 1 from 1.4). Applying the induction method we shall prove that all x_n belong to the interval $[a, b]$, to be more exact, on the interval $[c, b]$, where c is the desired root. Since $x_0 = b$ lies on the interval $[c, b]$, for the induction to be carried out it is sufficient to prove, under the assumption that x_n belongs to the interval $[c, b]$, that x_{n+1} also belongs to that interval. If $x_n = c$, then $f(x_n) = f(c) = 0$ and it follows from formula (12.9) that $x_{n+1} = x_n = c$, that is, we have carried out the induction. Suppose now that $x_n > c$. Then, taking into account that $f(c) = 0$, we obtain from formula (12.9)

$$x_n - x_{n+1} = \frac{f(x_n) - f(c)}{f'(x_n)}.$$

Applying Lagrange's formula to the expression appearing in the numerator of the last fraction, we find

$$x_n - x_{n+1} = (x_n - c)\frac{f'(\xi_n)}{f'(x_n)},$$

where $c < \xi_n < x_n$. Since the derivative is nondecreasing and positive, the fraction $\frac{f'(\xi_n)}{f'(x_n)}$ is positive and does not exceed unity, i.e. $x_n - x_{n+1} \leqslant x_n - c$ or $x_{n+1} \geqslant c$. Thus we have carried out the induction. Inasmuch as the derivative $f'(x)$ is positive, the function (x) increases and, therefore, the inequality $c \leqslant x_n$ yields $0 = f(c) \leqslant f(x_n)$. Thus we have that $\frac{f(x_n)}{f'(x_n)} \geqslant 0$. Hence, by virtue of formula (12.9), we have that $x_{n+1} \leqslant x_n$, that is, the sequence $\{x_n\}$ does not increase. Since, besides that, the sequence is bounded below by the number c, it converges (see Theorem 3.15). In accordance with Assertion 1 from 12.1.4, its limit is the desired root c.

Remark 3. We have considered the case where $f'(x)$ is nondecreasing and positive on $[a, b]$. Three other cases are possible: (1) $f'(x)$ is nonincreasing and negative on $[a, b]$, (2) $f'(x)$ is nonincreasing and positive on $[a, b]$, (3) $f'(x)$ is nondecreasing and negative on $[a, b]$.

In all the three cases the tangent method can be substantiated by a complete analogy with the case discussed above. We must only point out that in case (1) the value $x_0 = b$ should be taken as the zero approximation, while in cases (2) and (3) the value $x_0 = a$. This will guarantee the belonging of all the terms of the iterative sequence $\{x_n\}$ to the interval $[a, b]$ and the convergence of the sequence to the required root c.

Remark 4. We shall demonstrate the estimation of the deviation of the nth approximation x_n from the exact value of the root c (under the assumptions formulated in this subsection).

Let us apply Lagrange's formula to the expression $f(x_n) = f(x_n) - f(c)$. We shall have $f(x_n) = (x_n - c) f'(\xi_n)$. From this we obtain the following evaluation:

$$|x_n - c| \leqslant \frac{|f(x_n)|}{m}, \tag{12.13}$$

where m is the minimum value of $|f'(x)|$ on the interval $[a, b]$. Formula (12.13) makes it possible to estimate the deviation of x_n from the exact value of the root c in terms of the value of the modulus of the given function $y = f(x)$ at point x_n.

12.1.6. Substantiation of the chord method. Suppose that the required root c of the equation $f(x) = 0$ is isolated on a certain closed interval $[a, b]$ on which $f(x)$ possesses a *monotonic and continuous first derivative retaining a constant sign.* We shall assume, for definiteness, that this derivative does not decrease and is positive on the interval $[a, b]$. Note that the equation

$$x = F(x), \text{ where } F(x) = x - \frac{(b-x) f(x)}{f(b) - f(x)} \; * \tag{12.14}$$

possesses on the interval $[a, b]$ only one root c coinciding with the root of the equation $f(x) = 0$. Therefore, instead of the equation $f(x) = 0$ we shall try to solve equation (12.14). For that purpose, having fixed $x_0 = a$, we shall construct the iterative sequence

$$x_{n+1} = F(x_n) = x_n - \frac{(b-x_n) f(x_n)}{f(b) - f(x_n)}. \tag{12.15}$$

Note that the recurrence formula (12.15) coincides exactly with the recurrence formula (12.2).

Let us prove that the sequence $\{x_n\}$ converges to the required root c. For that purpose it is sufficient to prove that all x_n lie on the closed interval $[a, b]$ and that the sequence $\{x_n\}$ is convergent (see Assertion 1 from 12.1.4). Applying the induction method *we shall prove that all x_n lie on the interval $[a, b]$, to be more exact, on the interval $[a, c]$, where c is the required root.* Since the value $x_0 = a$ belongs to the interval $[a, c]$, to carry out the induction it is sufficient to assume that x_n lies on the interval $[a, c]$ and to prove that x_{n+1} also lies on that interval. Making use of formula (12.15) and taking into account that $f(c) = 0$, we shall have**

$$x_{n+1} - x_n = -\frac{(b-x_n) f(x_n)}{f(b) - f(x_n)} = \frac{(b-x_n)[f(c) - f(x_n)]}{[f(b) - f(c)] + [f(c) - f(x_n)]}.$$

---

* In this case we assume that $F(b) = b - \dfrac{f(b)}{f'(b)}$. Then the function $F(x)$ is continuous throughout the interval $[a, b]$.

** We assume in what follows that $x_n < c$, for if $x_n = c$, then $f(x_n) = f(c) = 0$ and, hence, $x_{n+1} = x_n = c$, i.e. it is an established fact that x_{n+1} belongs to the interval $[a, c]$.

Applying Lagrange's formula to the expressions in square brackets, we obtain

$$x_{n+1}-x_n=\frac{(b-x_n)\,f'(\xi_n)\,(c-x_n)}{(b-c)\,f'(\xi_n^*)+(c-x_n)f'(\xi_n)}, \tag{12.16}$$

where $x_n<\xi_n<c$, $c<\xi_n^*<b$, i.e. $\xi_n<\xi_n^*$. The derivative $f'(x)$ being nondecreasing and positive, we can write $0<f'(\xi_n)\leqslant f'(\xi_n^*)$, whence, since $b-c>0$ and $c-x_n>0$, we obtain

$$(b-c)\,f'(\xi_n^*)+(c-x_n)\,f'(\xi_n)\geqslant[(b-c)+(c-x_n)]\,f'(\xi_n)=$$

$$=(b-x_n)\,f'(\xi_n).$$

Thus we can find, from equality (12.16), that $x_{n+1}-x_n\leqslant c-x_n$, or $x_{n+1}\leqslant c$, that is, we have carried out the induction.

Let us prove now that the sequence $\{x_n\}$ is nondecreasing. To do that, it is sufficient to prove that the fraction appearing on the right-hand side of equality (12.15) is nonpositive. The derivative $f'(x)$ being positive on the closed interval $[a, b]$, the function $f(x)$ increases on that interval and, therefore, from the inequality $x_n\leqslant c<b$ it follows that $f(x_n)\leqslant f(c)=0$, $f(b)-f(x_n)>0$. Hence follows the nonpositivity of the indicated fraction.

Thus we see that the sequence $\{x_n\}$ does not decrease and is bounded above by the number c. By Theorem 3.15 this sequence is convergent. By virtue of Assertion 1 from 2.1.4 its limit is the root c.

Remark 1. We have considered the case where $f'(x)$ does not decrease and is positive on $[a, b]$. Three other cases are possible: (1) $f'(x)$ does not increase and is negative on $[a, b]$, (2) $f'(x)$ does not increase and is positive on $[a, b]$, (3) $f'(x)$ does not decrease and is negative on $[a, b]$.

These three cases are analogous to the case discussed above. In case (1) the equation $f(x)=0$ is replaced by equation (12.14) as above and $x_0=a$ is taken as the zero approximation (in this case the sequence $\{x_n\}$ turns out to be nondecreasing as well). In cases (2) and (3) the equation $f(x)=0$ is replaced not by equation (12.14) but by the equation

$$x=F(x),$$

where

$$F(x)=x-\frac{(a-x)\,f(x)}{f(a)-f(x)}$$

and the point $x_0=b$ is taken as the zero approximation (in this case the sequence $\{x_n\}$ turns out to be nonincreasing).

Remark 2. It should be pointed out that for the chord method the same evaluation (12.13) of the deviation of x_n from the root c is valid as for the method of tangents.

Remark 3. A combined method is often used in practice consisting in successive application of the method of chords and the tangent method. We shall assume for the sake of definiteness that $f'(x)$ is nondecreasing and positive on the closed interval $[a, b]$ (Fig. 12.6). Let us specify x_1 by the tangent method, taking point b as the zero approximation. Then we shall specify x_2 applying the chord method,

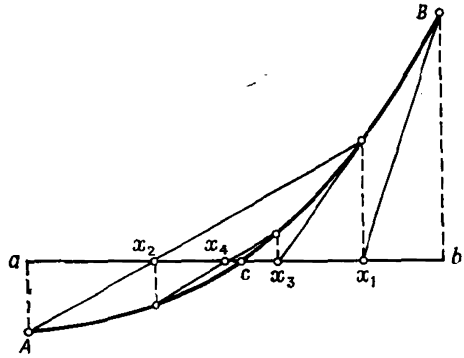


Fig. 12.6

however not to the interval $[a, b]$ but to the interval $[a, x_1]$. Next, proceeding from the known x_1 and x_4 obtained by the chord method we shall specify x_3 by the tangent method, applying it to the interval $[x_2, x_3]$. This process is illustrated in Fig. 12.6.

The advantages of the combined method are the following: first, it allows for a more rapid convergence than the chord method and, second, since the successive approximations x_n and x_{n+1} of the combined method approximate the root from both sides, the difference $|x_{n+1}-x_n|$ gives the evaluation of the error of this method. If we take $x_n^*=\frac{x_n+x_{n+1}}{2}$ for an approximate value of the root, we shall obtain the following evaluation of the error:

$$|x_n^*-c|<\frac{|x_{n+1}-x_n|}{2}.$$

12.2. APPROXIMATE METHODS OF CALCULATING DEFINITE INTEGRALS

12.2.1. Introductory remarks. When solving some actual physical and technological problems we sometimes come across definite integrals of functions whose antiderivatives *cannot be expressed in terms of elementary functions.* Besides, we sometimes meet in applications definite integrals *whose integrand functions themselves are not*

elementary. Thus a necessity arises to elaborate approximate methods of calculating definite integrals*.

In this subsection we shall get acquainted with three approximate methods most often used for computing definite integrals: *the method of rectangles, the method of trapezoids and the method of parabolas.*

The main idea underlying these methods is the replacement of the integrand function $f(x)$ by a function of a more simple nature,

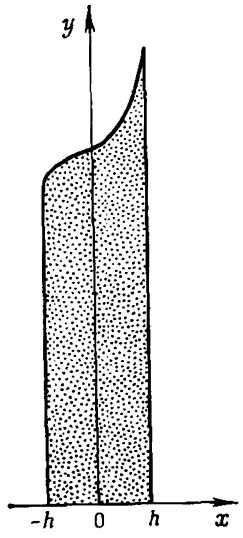


Fig. 12.7

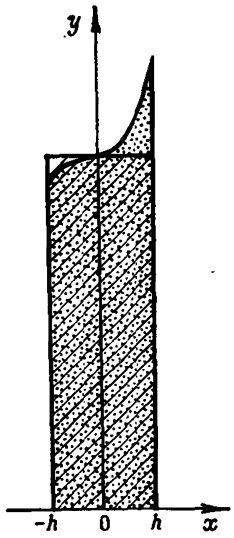


Fig. 12.8

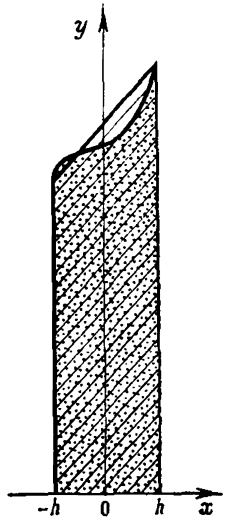


Fig. 12.9

a polynomial coinciding with $f(x)$ at some points. To elucidate this idea, let us consider, at small h, the integral $\int_{-h}^{h} f(x)\,dx$ which is the area of a narrow curvilinear trapezoid located under the graph of the function $y = f(x)$ on the closed interval $[-h, h]$ (Fig. 12.7).

Let us replace the function $f(x)$ by a polynomial of order zero, namely, by the constant $f(0)$. As a result, the integral $\int_{-h}^{h} f(x)\,dx$ will be approximately replaced by the *area of the rectangle* hatched in Fig. 12.8. We shall show below that with certain requirements imposed on $f(x)$ the error resulting from the replacement is of the order of h^3. Let us now replace the function $f(x)$ by a polynomial of the first order, namely, by the linear function $y = kx + b$ coinciding

* Note that approximate methods are also often used for integrals which can be expressed in terms of elementary functions.

with $f(x)$ at points $-h$ and h. As a result, the integral $\int_{-h}^{h} f(x)\,dx$ will be approximately replaced by the *area of the rectilinear trapezoid* hatched in Fig. 12.9. We shall show below that with certain requirements imposed on $f(x)$ the error resulting from this replacement is also of the order of h^3. Let us finally replace the function $f(x)$ by a polynomial of the second order, that is, by the parabola $y = Ax^2 + Bx + C$, coinciding with $f(x)$ at points $-h$, 0 and h. In this case the integral $\int_{-h}^{h} f(x)\,dx$ will be approximately replaced by the *area of the figure located under the parabola* and hatched in Fig. 12.10.

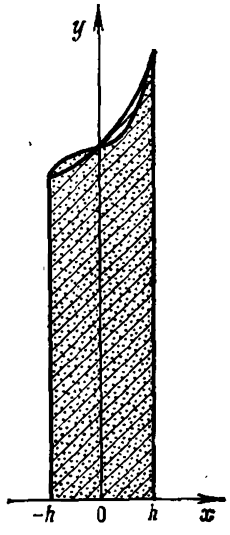


Fig. 12.10

We shall show below that when certain requirements are imposed on the function $f(x)$ the error committed in the replacement is of the order of h^5. If we have to calculate the integral $\int_{a}^{b} f(x)\,dx$ over any closed interval $[a, b]$, it is natural to divide the interval into a sufficiently large number of small closed intervals and apply the reasoning presented above to each of these intervals. We shall thus arrive at the methods of rectangles, trapezoids and parabolas in their general form.

The detailed exposition of these three methods will be given below. Here we shall only make a remark significant for our further argument.

Remark. *Assume that the function $f(x)$ is continuous on the closed interval $[a, b]$ and $x_1, x_2, \ldots, x_n$ are some points belonging to the interval $[a, b]$. Then there is a point ξ on that interval such that the arithmetic mean $\frac{f(x_1)+f(x_2)+\cdots+f(x_n)}{n}$ is equal to $f(\xi)$.*

In fact, let us designate by m and M the infimum and the supremum of the function $f(x)$ on the interval $[a, b]$. Then the inequalities $m \leqslant f(x_k) \leqslant M$ $(k = 1, 2, \ldots, n)$ are valid for any number k. Summing up these inequalities over all the numbers $k = 1, 2, \ldots, n$ and dividing the result by n, we obtain

$$m \leqslant \frac{f(x_1)+f(x_2)+\ldots+f(x_n)}{n} \leqslant M.$$

Since a continuous function can assume any intermediate value contained between m and M, there can be found a point ξ on the interval

$[a, b]$ such that

$$f(\xi)=\frac{f(x_1)+f(x_2)+\ldots+f(x_n)}{n}. \tag{12.17}$$

12.2.2. The method of rectangles. Suppose we have to calculate the integral

$$\int_a^b f(x)\,dx. \tag{12.18}$$

Let us partition the interval $[a, b]$ into n *equal* parts by the points $a=x_0<x_2<\ldots<x_{2n}=b$ and designate the middle point of the

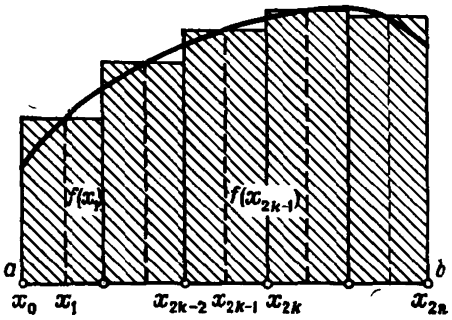


Fig. 12.11

interval $[x_{2k-2}, x_{2k}]$ by x_{2k-1} (Fig. 12.11). The method of rectangles consists in the replacement of integral (12.18) by the sum

$$\frac{b-a}{n}[f(x_1)+f(x_3)+\ldots+f(x_{2n-1})]$$

of the areas of the rectangles with altitudes equal respectively to $f(x_{2k-1})$ and bases equal to $x_{2k}-x_{2k-2}=(b-a)/n$ (these reactangles are hatched in Fig. 12.11). Thus there holds a formula

$$\int_a^b f(x)\,dx=\frac{b-a}{n}[f(x_1)+f(x_3)+\ldots+f(x_{2n-1})]+R, \tag{12.19}$$

where R is the remainder. Equation (12.19) is called the *formula of rectangles.*

Let us prove that if the function $f(x)$ possesses a continuous second derivative on the closed interval $[a, b]$, then there is a point η on that interval such that the remainder R in formula (12.19) is equal to

$$R=\frac{(b-a)^3}{24n^2}f^{(2)}(\eta). \tag{12.20}$$

With that aim in view we shall first estimate $\int_{-h}^{h} f(x)\,dx$ assuming that *the function* $f(x)$ *possesses a continuous second derivative on the closed interval* $[-h, h]$.

To do that, let us twice integrate by parts each of the following two integrals:

$$I_1 = \int_{-h}^{0} f^{(2)}(x)\,(x+h)^2\,dx, \quad I_2 = \int_{0}^{h} f^{(2)}(x)\,(x-h)^2\,dx.$$

For the first of these integrals we obtain

$$I_1 = \int_{-h}^{0} f^{(2)}(x)\,(x+h)^2\,dx = [(x+h)^2 f'(x)]\Big|_{-h}^{0} -$$

$$-2\int_{-h}^{0} f'(x)\,(x+h)\,dx = f'(0)\,h^2 - [2(x+h)\,f(x)]\Big|_{-h}^{0} +$$

$$+2\int_{-h}^{0} f(x)\,dx = f'(0)\,h^2 - 2f(0)\,h + 2\int_{-h}^{0} f(x)\,dx.$$

By a complete analogy we obtain for the second integral

$$I_2 = -f'(0)\,h^2 - 2f(0)\,h + 2\int_{0}^{h} f(x)\,dx.$$

The half-sum of the expressions obtained for I_1 and I_2 leads to the following formula:

$$\int_{-h}^{h} f(x)\,dx = 2f(0)\,h + \frac{I_1+I_2}{2}. \tag{12.21}$$

Let us evaluate the quantity $(I_1 + I_2)/2$ applying the mean-value formula to the integrals I_1 and I_2 and taking into account the nonnegativity of the functions $(x+h)^2$ and $(x-h)^2$. We shall see that there is a point ξ_1 on the closed interval $[-h, 0]$ and a point ξ_2 on the closed interval $[0, h]$ such that

$$\frac{I_1+I_2}{2} = \frac{1}{2}\int_{-h}^{0} f^{(2)}(x)\,(x+h)^2\,dx + \frac{1}{2}\int_{0}^{h} f^{(2)}(x)\,(x-h^2)\,dx =$$

$$= \frac{1}{2} f^{(2)}(\xi_1)\int_{-h}^{0} (x+h)^2\,dx + \frac{1}{2} f^{(2)}(\xi_2)\int_{0}^{h} (x-h)^2\,dx =$$

$$= \frac{h^3}{6} f^{(2)}(\xi_1) + \frac{h^3}{6} f^{(2)}(\xi_2) = \frac{h^3}{3}\,\frac{f^{(2)}(\xi_1) + f^{(2)}(\xi_2)}{2}.$$

By virtue of the remark made at the end of 2.1, there is a point η on the interval $[-h, h]$ such that

$$\frac{f^{(2)}(\xi_1)+f^{(2)}(\xi_2)}{2}+f^{(2)}(\eta).$$

Therefore, we shall have the following expression for the half-sum $\frac{I_1+I_2}{2}$:

$$\frac{I_1+I_2}{2}=\frac{h^3}{3}f^{(2)}(\eta).$$

Substituting this expression into (12.21), we obtain

$$\int_{-h}^{h} f(x)\,dx=2f(0)\,h+\overline{R}, \tag{12.22}$$

where

$$\overline{R}=\frac{1}{24}f^{(2)}(\eta)\,(2h)^3 \quad (-h\leqslant\eta\leqslant h). \tag{12.23}$$

Since the quantity $2f(0)\,h$ is the area of the rectangle hatched in Fig. 12.8, formulas (12.22) and (12.23) prove that the error committed in the replacement of $\int_{-h}^{h} f(x)\,dx$ by the indicated area is of the order of h^3.

Thus, *the smaller the quantity h, the more exact is the formula* $\int_{-h}^{h} f(x)\,dx \approx 2f(0)\,h$. Therefore, to calculate the integral $\int_a^b f(x)\,dx$, it is natural to represent it as the sum of a sufficiently large number n of the integrals

$$\int_{x_0}^{x_2} f(x)\,dx+\int_{x_2}^{x_4} f(x)\,dx+\ldots+\int_{x_{2n-2}}^{x_{2n}} f(x)\,dx$$

and to apply formula (12.22) to each of the indicated integrals. Also taking into account that the length of the interval $[x_{2k-2}, x_{2k}]$ is equal to $(b-a)/n$, we shall obtain the formula for rectangles (12.19) in which

$$R=\overline{R}_1+\overline{R}_2+\ldots+\overline{R}_n=\frac{(b-a)^3}{24n^3}[f^{(2)}(\eta_1)+f^{(2)}(\eta_2)+\ldots$$

$$\ldots+f^{(2)}(\eta_n)]=\frac{(b-a)^3}{24n^2}\,\frac{f^{(2)}(\eta_1)+f^{(2)}(\eta_2)+\ldots+f^{(2)}(\eta_n)}{n}=$$

$$=\frac{(b-a)^3}{24n^2}f^{(2)}(\eta).$$

(Here $a \leqslant \eta \leqslant b$. We have used formula (12.17) for the function $f^{(2)}(x)$.)

12.2.3. The method of trapezoids. Suppose that, as before, we have to calculate the integral

$$\int_a^b f(x)\,dx. \tag{12.18}$$

Let us divide the closed interval $[a, b]$ into n equal parts by means

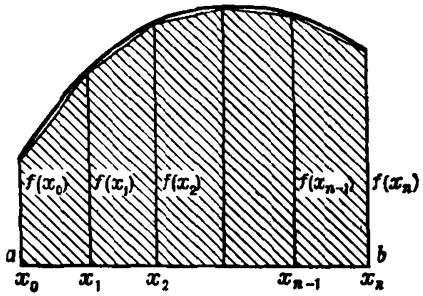


Fig. 12.12

of the points $a = x_0 < x_1 < x_2 < \ldots < x_n = b$ (Fig. 12.12). The method of trapezoids consists in the replacement of integral (12.18) by the sum

$$\frac{b-a}{2n}\{[f(x_0)+f(x_1)]+[f(x_1)+f(x_2)]+\ldots+[f(x_{n-1})+f(x_n)]\} =$$

$$=\frac{b-a}{2n}\left\{f(a)+f(b)+2\sum_{k=1}^{n-1}f(x_k)\right\}$$

of the areas of the trapezoids with the bases equal respectively to $f(x_{k-1})$ and $f(x_k)$ and with altitudes equal to $x_k - x_{k-1} = (b-a)/n$ (these trapezoids are hatched in Fig. 12.12).

Thus, there holds a formula

$$\int_a^b f(x)\,dx = \frac{b-a}{2n}\left\{f(a)+f(b)+2\sum_{k=1}^{n-1}f(x_k)\right\}+R, \tag{12.24}$$

where R is the remainder. Equation (12.24) is known as the *formula of trapezoids.*

Let us prove that if the function $f(x)$ possesses a continuous second derivative on the closed interval $[a, b]$, then there is a point η on that interval such that the remainder R in formula (12.24) is of the

form

$$R = -\frac{(b-a)^3}{12n^2} f^{(2)}(\eta). \tag{12.25}$$

We shall first evaluate the integral $\int_{-h}^{h} f(x)\,dx$, assuming that *the function $f(x)$ possesses a continuous second derivative on the interval* $[-h, h]$.

Twice integrating the integral $\int_{-h}^{h} f^{(2)}(x)(x^2 - h^2)\,dx$ by parts, we obtain

$$\int_{-h}^{h} f^{(2)}(x)(x^2-h^2)\,dx = [f'(x)(x^2-h^2)]\Big|_{-h}^{h} -$$

$$-2\int_{-h}^{h} f'(x)\,x\,dx = -[2f(x)\,x]\Big|_{-h}^{h} + 2\int_{-h}^{h} f(x)\,dx =$$

$$= -2[f(-h)+f(h)]h + 2\int_{-h}^{h} f(x)\,dx. \tag{12.26}$$

By virtue of (12.26) we arrive at the formula

$$\int_{-h}^{h} f(x)\,dx = \frac{f(-h)+f(h)}{2}\,2h + \bar{R}, \tag{12.27}$$

where

$$\bar{R} = -\frac{1}{12} f^{(2)}(\eta)(2h)^3 \quad (-h \leqslant \eta \leqslant h). \tag{12.28}$$

Since the quantity $\frac{f(-h)+f(h)}{2}\,2h$ is the area of the trapezoid hatched in Fig. 12.9, formulas (12.27) and (12.28) prove that the error we commit by replacing $\int_{-h}^{h} f(x)\,dx$ by the indicated area is of the order of h^3.

To calculate the integral $\int_{a}^{b} f(x)\,dx$, we represent this integral in the same way as in the method of rectangles, i.e. as the sum of a sufficiently large number n of the integrals

$$\int_{x_0}^{x_1} f(x)\,dx + \int_{x_1}^{x_2} f(x)\,dx + \ldots + \int_{x_{n-1}}^{x_n} f(x)\,dx.$$

Applying formulas (12.27) and (12.28) to each of these integrals, we arrive at the trapezoid formula (12.24) with expression (12.25) for the remainder.

12.2.4. The method of parabolas. To calculate the integral

$$\int_a^b f(x)\,dx \tag{12.18}$$

let us divide once again the closed interval $[a, b]$ into n equal parts by means of the points $a = x_0 < x_2 < \ldots < x_{2n} = b$ and designate the middle point of the interval $[x_{2k-2}, x_{2k}]$ by x_{2k-1}. The method of parabolas consists in replacing integral (12.18) by the sum

$$\frac{b-a}{6n}\{[f(x_0)+4f(x_1)+f(x_2)]+[f(x_2)+4f(x_3)+f(x_4)]+\ldots$$
$$\ldots+[f(x_{2n-2})+4f(x_{2n-1})+f(x_{2n})]\}=$$
$$=\frac{b-a}{6n}\left\{f(a)+f(b)+2\sum_{k=1}^{n-1}f(x_{2k})+4\sum_{k=0}^{n-1}f(x_{2k+1})\right\}$$

of the areas of the figures hatched in Fig. 12.13 and constituting

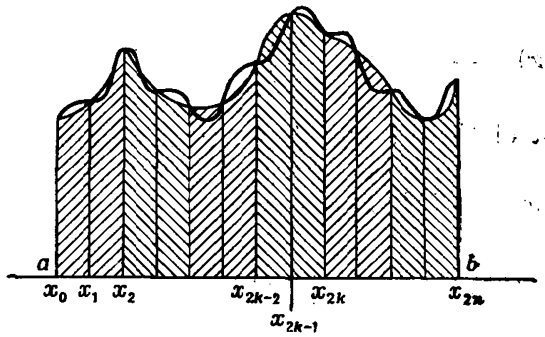


Fig. 12.13

curvilinear trapezoids located under the parabolas passing through three points of the graph of the function $f(x)$ with the abscissas x_{2k-2}, x_{2k-1} and x_{2k}*.

* It follows from Example 2 cited in 11.2.4 that with due account of the fact that $\frac{b-a}{6n}=\frac{x_{2k}-x_{2k-2}}{6}$, the expression $\frac{b-a}{6n}[f(x_{2k-2})+4f\times$ $\times(x_{2k-1})+f(x_{2k})]$ is the area located under the parabola which passes through three points of the graph of the function $f(x)$ with the abscissas x_{2k-2}, x_{2k-1} and x_{2k}.

Thus, the following formula holds true:

$$\int_a^b f(x)\,dx = \frac{b-a}{6n}\left[f(a)+f(b)+2\sum_{k=1}^{n-1} f(x_{2k})+4\sum_{k=0}^{n-1} f_1'(x_{2k+1})\right]+R, \tag{12.29}$$

where R is a remainder. Equation (12.29) is called the *formula of parabolas* or *Simpson's* rule.*

Let us prove that if the function $f(x)$ possesses a continuous fourth derivative on the closed interval $[a, b]$ then there is a point η on that interval such that the remainder R in formula (12.29) is

$$R = -\frac{(b-a)^5}{2880n^4} f^{(4)}(\eta). \tag{12.30}$$

With this aim in view, we shall first estimate $\int_{-h}^{h} f(x)\,dx$, assuming that *the function $f(x)$ possesses a continuous fourth derivative on the closed interval $[-h, h]$.*

For that purpose we shall integrate by parts the following two integrals four times:

$$I_1 = \int_{-h}^{0} f^{(4)}(x)(x+h)^3\left(x-\frac{h}{3}\right)dx,$$

$$I_2 = \int_{0}^{h} f^{(4)}(x)(x-h)^3\left(x+\frac{h}{3}\right)dx.$$

For the first of these integrals we shall obtain

$$\begin{aligned} I_1 &= \int_{-h}^{0} f^{(4)}(x)(x+h)^3\left(x-\frac{h}{3}\right)dx = \\ &= \left[f^{(3)}(x)(x+h)^3\left(x-\frac{h}{3}\right)\right]\Big|_{-h}^{0} - \\ &- \left\{f^{(2)}(x)\left[3(x+h)^2\left(x-\frac{h}{3}\right)+(x+h)^3\right]\right\}\Big|_{-h}^{0} + \\ &+ \left\{6f'(x)\left[(x+h)\left(x-\frac{h}{3}\right)+(x+h)^2\right]\right\}\Big|_{-h}^{0} - \\ &- \left\{6f(x)\left(4x+\frac{8}{3}h\right)\right\}\Big|_{-h}^{0} + 24\int_{-h}^{0} f(x)\,dx = -f^{(3)}(0)\frac{h^4}{3} + \\ &+ 4f'(0)h^2 - 8h[f(-h)+2f(0)] + 24\int_{-h}^{0} f(x)\,dx. \end{aligned} \tag{12.31}$$

* Thomas Simpson (1710-1761), a British mathematician.

By a complete analogy we obtain for I_2

$$I_2 = f^{(3)}(0)\frac{h^4}{3} - 4f'(0)\,h^2 - 8h\,[f(h) + 2f(0)] + 24\int_0^h f(x)\,dx. \tag{12.32}$$

By adding up relations (12.31) and (12.32) we obtain the following equality:

$$\int_{-h}^{h} f(x)\,dx = \frac{f(-h)+4f(0)+f(h)}{6}\,2h + \frac{I_1+I_2}{24}. \tag{12.33}$$

To evaluate $\frac{I_1+I_2}{24}$ we shall apply the mean-value formula to the integrals I_1 and I_2, taking into consideration the non-positivity of the functions $(x+h)^3\left(x-\frac{h}{3}\right)$ and $(x-h)^3\left(x+\frac{h}{3}\right)$ on the closed intervals $[-h, 0]$ and $[0, h]$ respectively.

We shall see that there is a point ξ_1 on the interval $[-h, 0]$ and a point ξ_2 on the interval $[0, h]$ such that

$$\frac{I_1+I_2}{24} = \frac{1}{24}\left[f^{(4)}(\xi_1)\int_{-h}^{0}(x+h)^3\left(x-\frac{h}{3}\right)dx + \right.$$

$$\left. + f^{(4)}(\xi_2)\int_0^h (x-h)^3\left(x+\frac{h}{3}\right)dx\right] = \frac{-h^5}{90}\,\frac{f^{(4)}(\xi_1)+f^{(4)}(\xi_2)}{2}.$$

Using again the remark made at the end of 2.1, we see that there is a point η on the closed interval $[-h, h]$ such that

$$\frac{I_1+I_2}{24} = \frac{-h^5}{90}\,f^{(4)}(\eta). \tag{12.34}$$

From (12.33) and (12.34) we finally obtain

$$\int_{-h}^{h} f(x)\,dx = \frac{[f(-h)+4f(0)+f(h)]}{6}\,2h + \bar{R}, \tag{12.35}$$

where

$$\bar{R} = -\frac{(2h)^5}{2880}\,f^{(4)}(\eta). \tag{12.36}$$

Since the quantity $\frac{[f(-h)+4f(0)+f(h)]}{6}\,2h$ is the area of the figure lying under the parabola and hatched in Fig. 12.10, formulas (12.35) and (12.36) prove that the error we commit by replacing $\int_{-h}^{h} f(x)\,dx$ by the indicated area is of the order of h^5.

In the same manner as in the methods of rectangles and trapezoids, in order to calculate the integral $\int_a^b f(x)\,dx$ we shall represent it as the sum of n integrals

$$\int_{x_0}^{x_2} f(x)\,dx+\int_{x_2}^{x_4} f(x)\,dx+\ldots+\int_{x_{2n-2}}^{x_{2n}} f(x)\,dx.$$

Applying formulas (12.35) and (12.36) to each of these integrals, we arrive at Simpson's formula (12.29) with expression (12.30) for the remainder.

Comparing the remainder (12.30) with the remainders (12.20) and (12.25), we make sure that Simpson's formula yields more exact solutions than the formulas of rectangles and trapezoids.

To illustrate the application of Simpson's formula, let us calculate the integral $I(x_0)=\int_0^{x_0} e^{-x^2}dx$*, restricting ourselves, for simplicity, by the values x_0 from the closed interval $0\leqslant x_0\leqslant 1$. Fixing $f(x)=e^{-x^2}$ and calculating the derivative $f^{(4)}(x)=4(4x^4-12x^2+3)e^{-x^2}$, we shall easily ascertain the fact that for all x from the closed interval $0\leqslant x\leqslant x_0\leqslant 1$ $|f^{(4)}(x)|<20$ in all the cases. Proceeding from the evaluation (12.30), we can assert that $|R|<\frac{1}{144n^4}$. Thus we see that by dividing the interval $[0, x_0]$ only into five equal parts and by replacing the integral under consideration by the sum appearing on the right-hand side of Simpson's formula, we shall calculate this integral with an accuracy to within $\frac{1}{144\cdot 5^4}=\frac{1}{90\,000}$.

12.2.5. Concluding remarks. Each of the methods of calculating roots of equations and definite integrals presented in this chapter *contains a strictly formulated algorithm* for carrying out calculations. Another feature of the methods we have considered is the *similitude* of the computing operations which must be performed at each stage. These two features ensure the possibility of wide application of the methods having been discussed for carrying out calculations on modern computers.

Somewhat earlier, in order to carry out approximate calculations of integral (12.18) of the function $f(x)$, we have proceeded from partitioning of the main interval $[a, b]$ into a sufficiently large number n of *equal* subintervals of the same length h and from the consequent

* The integral in question cannot be expressed in terms of elementary functions. This integral finds wide application in statistical physics, in the theory of heat conductivity and in diffusion.

replacement of the function $f(x)$ on each subinterval by a polynomial of the zeroth, first and second order respectively.

An error resulting from such an approach does not take into account the individual features of the function $f(x)$. The idea naturally arises as to the variation of the points of partitioning the main interval $[a, b]$ and as to the choice, for every fixed function $f(x)$, of the optimal division of the main interval $[a, b]$ into n nonequal, in general, subintervals such that would ensure the minimum value of the error of the given approximate formula.

In the supplement to Chap. 14 we shall deal with the realization of this idea belonging to A. N. Tikhonov and S. S. Gaisaryan.

CHAPTER 13

THEORY OF NUMBER SERIES

As early as when studying the elementary course, we came across sums consisting of an *infinite* number of summands (say, the sums of an infinite number of terms of a geometric progression). Sums of this kind, called *series*, are the subject of the present chapter. We shall ascertain that under certain conditions series possess properties analogous to the properties of finite sums.

13.1. THE CONCEPT OF A NUMBER SERIES

13.1.1. A series and its partial sums. Convergent and divergent series. Let us consider an infinite number sequence u_1, u_2,..., ..., u_k, ... and formally form from the terms of this sequence an expression

$$u_1 + u_2 + \ldots + u_k + \ldots = \sum_{k=1}^{\infty} u_k. \tag{13.1}$$

It is customary to call an expression of the form of (13.1) a *number series* or simply a *series*. Separate elements u_k constituting expression (13.1) are usually called the *terms of the series*. We shall accept the symbol of the sum $\sum$ to designate the series.

The sum of the first n terms of a given series will be called an n-th partial sum *and denoted by* S_n. Thus we have $S_n = u_1 + u_2 + \ldots + u_n = \sum_{k=1}^{n} u_k$. *Series* (13.1) *is called* **convergent** *if the sequence* $\{S_n\}$ *of the partial sums of that series is convergent. In that case the limit S of the sequence of the partial sums* $\{S_n\}$ *is called the* **sum** *of the given series.* Thus, for a convergent series having the sum S we can formally write an equality

$$S = \sum_{k=1}^{\infty} u_k.$$

In the case where $\lim_{n \to \infty} S_n$ *does not exist, the series is called* **divergent.**

It should be emphasized that the concept of a sum is determined only for a covergent series and, as distinct from the concept of a finite sum, is introduced by means of a limiting process*.

It should be pointed out that the consideration of number series is a new form of investigation of number sequences, for (1) each given series is unambiguously associated with the sequence of its partial sums, (2) each given sequence $\{S_n\}$ is unambiguously associated with a series for which this sequence is the sequence of its partial sums (it is sufficient to put the terms of the series equal to $u_k = S_k - S_{k-1}$ for $k > 1$ and $u_1 = S_1$).

One of the principal problems of the theory of number series is the determination of the tests for convergence or divergence of a given series.

Examples of number series.

1. Let us decide whether the following series is convergent:

$$1-1+1-1+\ldots=\sum_{k=1}^{\infty}(-1)^{k-1}. \tag{13.2}$$

Since the sequence of its partial sums $S_1 = 1, S_2 = 0, \ldots$ $\ldots, S_{2n-1} = 1, S_{2n} = 0, \ldots$ is not limited, series (13.2) is divergent.

2. Let us now consider a series consisting of the terms of a geometric progression:

$$1+q+q^2+\ldots+q^k+\ldots=\sum_{k=1}^{\infty}q^{k-1}, \tag{13.3}$$

whose n-th partial sum S_n, for $q \neq 1$, has the form

$$S_n = 1+q+\ldots+q^{n-1}=\frac{1-q^n}{1-q}=\frac{1}{1-q}-\frac{q^n}{1-q}. \tag{13.4}$$

It is evident that for $|q| < 1$ the sequence of the partial sums S_n converges and has a limit equal to $1/(1-q)$. Thus we see that for $|q| < 1$ the series in question converges and possesses a sum equal to $1/(1-q)$.

For $|q| > 1$, it can be seen from equality (13.4) that the sequence S_n (and, hence, the series being considered) is divergent. For $|q| = 1$, the divergence of series (13.3) can be immediately seen. Indeed, for $q = +1$, $S_n = n$, the divergence of the sequence S_n is evident, and for $q = -1$ series (13.3) passes into series (13.2) which has been investigated.

* In modern mathematics, besides the concept of the sum indicated above, the concept of the sum of a series in various general meanings is also introduced. This makes it possible to sum up many divergent series in general senses (see Supplement 3 to the present chapter).

3. Suppose x is any fixed number. Let us prove that the series

$$1+\frac{x}{1!}+\frac{x^2}{2!}+\ldots+\frac{x^k}{k!}+\ldots=\sum_{k=1}^{\infty}\frac{x^{k-1}}{(k-1)!}\ * \tag{13.5}$$

converges and possesses a sum equal to e^x.

In 8.15.2 we obtained an expansion of the function e^x by Maclaurin's formula:

$$e^x=1+\frac{x}{1!}+\frac{x^2}{2!}+\ldots+\frac{x^{n-1}}{(n-1)!}+R_n(x), \tag{13.6}$$

where

$$R_n(x)=\frac{x^n}{n!}e^{\theta x}\quad(0<\theta<1). \tag{13.7}$$

From formulas (13.6) and (13.7) we obtain

$$\left|\left[1+\frac{x}{1!}+\frac{x^2}{2!}+\ldots+\frac{x^{n-1}}{(n-1)!}\right]-e^x\right|\leqslant\frac{|x|^n}{n!}e^{|x|}. \tag{13.8}$$

Denoting the nth partial sum of series (13.5) by S_n, we can rewrite inequality (13.8) in the form

$$|S_n-e^x|\leqslant\frac{|x|^n}{n!}e^{|x|}. \tag{13.9}$$

Since for any fixed x

$$\lim_{n\to\infty}\frac{|x|^n}{n!}=0\ **,$$

the right-hand side of inequality (13.9) is a term of an infinitesimal sequence. But that precisely means that the sequence $\{S_n\}$ *converges to the number* e^x. Consequently, series (13.5) converges and its sum is e^x.

4. By a complete analogy, using Maclaurin's formula for $\sin x$ and $\cos x$, we can prove that the series

$$x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\ldots=\sum_{k=1}^{\infty}\frac{(-1)^{k-1}x^{2k-1}}{(2k-1)!}$$

and

$$1-\frac{x^2}{2!}+\frac{x^4}{4!}-\frac{x^6}{6!}+\ldots=\sum_{k=1}^{\infty}\frac{(-1)^{k-1}x^{2k-2}}{(2k-2)!}$$

converge for any fixed value of x and possess sums equal to $\sin x$ and $\cos x$ respectively. (We suggest that the reader should verify this himself.)

* The symbol 0! is used to denote the number 1.

** See Example 3 in 3.3.3.

13.1.2. Cauchy's test for convergence of a series. Since, in accordance with the definition, the question concerning the convergence of a series is equivalent to that concerning the convergence of the sequence of its partial sums, we shall obtain the necessary and sufficient condition of convergence of the given series by formulating the Cauchy test for convergence of the sequence of its partial sums. For the sake of convenience, we shall present here the formulation of the Cauchy test for a sequence. *For the sequence* $\{S_n\}$ *to be convergent, it is necessary and sufficient that for any positive number* ε *there would be a number* N *such that for all the numbers* n *satisfying the condition* $n \geqslant N$ *and for all natural* p $(p = 1, 2, 3, \ldots)$

$$| S_{n+p} - S_n | < \varepsilon.$$

As a consequence of this assertion we obtain the following *fundamental* theorem.

Theorem 13.1. **(*Cauchy's test for a series.*)** *For the series* $\sum_{k=1}^{\infty} u_k$ *to converge, it is necessary and sufficient that for any positive number* ε *there should be a number* N *such that for all the numbers* n *complying with the condition* $n \geqslant N$ *and for all natural numbers* p

$$\left| \sum_{k=n+1}^{n+p} u_k \right| < \varepsilon. \tag{13.10}$$

To prove this theorem, it is sufficient to note that the quantity under the modulus sign in inequality (13.10) is equal to the difference of the partial sums $S_{n+p} - S_n$. It should be emphasized that Cauchy's test for convergence is only of theoretical interest. Its use for practical establishment of convergence or divergence of specific series usually evolves difficulties. Therefore, compliance with Cauchy's test does not obviate the necessity of establishing other practical effective criteria of convergence or divergence of series.

Two elementary but important corollaries can be derived from Theorem 13.1.

Corollary 1. *If the series* $\sum_{k=1}^{\infty} u_k$ *is convergent, then the sequence* $r_n = \sum_{k=n+1}^{\infty} u_k$ *is infinitesimal.* It is customary to call the quantity r_n the nth **remainder** *of the series* $\sum_{k=1}^{\infty} u_k$. To prove Corollary 1, it is sufficient to show that for any $\varepsilon > 0$ there is a number N such that $|r_n| \leqslant \varepsilon$ at $n \geqslant N$. The last inequality follows immediately from inequality (13.10) valid for *any* $p = 1, 2, 3, \ldots$, and from Theorem 3.13.

Corollary 2 (necessary condition of convergence of a series). *For the series* $\sum_{k=1}^{\infty} u_k$ *to be convergent it is necessary that the sequence* $u_1, u_2, u_3, \ldots$ *of the terms of that series should be infinitesimal.*

It is sufficient to prove that for the given convergent series and for any $\varepsilon > 0$ there is a number N_0 such that $|u_n| < \varepsilon$ at $n \geqslant N_0$. Suppose we are given some $\varepsilon > 0$. In accordance with Theorem 13.1 there is a number N such that inequality (13.10) holds for $n \geqslant N$ and for *any* natural p. In particular, for $p = 1$ this inequality has the form

$$|u_{n+1}| < \varepsilon \text{ (at } n \geqslant N). \tag{13.11}$$

If we now put the number N_0 equal to $N_0 = N + 1$, then we obtain $|u_n| < \varepsilon$ at $n \geqslant N_0$, by virtue of inequality (13.11), and that is what we had to prove.

Corollary 2 can also be formulated as follows: *for the convergence of the series* $\sum_{k=1}^{\infty} u_k$ *it is necessary to have* $\lim_{k\to\infty} u_k = 0$. Thus, when we investigate the convergence of a given series, we must find first of all whether the kth member of the series tends to zero as $k \to \infty$. If it does not, the series obviously diverges. Thus, for instance, the series

$$\sum_{k=1}^{\infty} \frac{k^2}{5k^2+300k}$$

is obviously divergent since

$$\lim_{k\to\infty} u_k = \lim_{k\to\infty} \frac{k^2}{5k^2+300k} = \frac{1}{5} \neq 0.$$

Quite similarly, the divergence of the series $\sum_{k=1}^{\infty} (-1)^{k-1}$ we have investigated follows from the fact that $\lim_{k\to\infty} (-1)^k$ does not exist.

It should be emphasized, however, that the tending of the kth term to zero for $k \to \infty$ is *only the necessary but not the sufficient condition of the convergence of the series.*

Let us take the following series as an example:

$$\sum_{k=1}^{\infty} \frac{1}{k} = 1 + \frac{1}{2} + \frac{1}{3} + \ldots \tag{13.12}$$

Such a series is usually called *harmonic*. It is evident that the necessary condition of convergence is complied with for the harmonic series

since $\lim\limits_{k\to\infty}\frac{1}{k}=0$. We shall prove, however, that the series is divergent. We shall make use of the Cauchy test and prove that for the positive number $\varepsilon = 1/2$ *there is no number* N such that at $n \geqslant N$, for any natural p,

$$\left|\sum_{k=n+1}^{n+p}\frac{1}{k}\right|<\varepsilon=\frac{1}{2}. \tag{13.13}$$

Indeed, if we take $p = n$, then *for arbitrarily large* n

$$\sum_{k=n+1}^{n+p}\frac{1}{k}=\sum_{k=n+1}^{2n}\frac{1}{k}\geqslant\frac{1}{2n}n=\frac{1}{2}.$$

(We have taken into account that there are n terms in the last sum and that the least of the summands is equal to $1/(2n)$.)

Thus we see that inequality (13.13) is not satisfied however large we take the number N. By virtue of Cauchy's test, series (13.12) diverges.

13.1.3. Two properties connected with the convergence of a series. 1°. *Deletion of a finite number of terms of a series (or addition of a finite number of terms to a series) does not affect its convergence or divergence.*

To ascertain this fact, it is sufficient to note that as a result of the deletion (or addition) of the terms, all the partial sums of the series, from a certain number onward, change by one and the same constant quantity.

2°. *If* c *is a nonzero constant,* $u'_k=cu_k$, *then the series* $\sum\limits_{k=1}^{\infty}u'_k$ *converges if and only if the series* $\sum\limits_{k=1}^{\infty}u_k$ *is convergent.*

If we designate by S'_n and S_n, respectively, the nth partial sums of the investigated series, then, evidently, $S'_n = cS_n$. From the last equality it follows that $\lim\limits_{n\to\infty}S'_n$ exists if and only if there exists $\lim\limits_{n\to\infty}S_n$.

13.2. SERIES WITH POSITIVE TERMS

13.2.1. A necessary and sufficient condition for convergence of a series with positive terms. We shall consider here series *whose all terms are nonnegative.* As is customary, we shall call such series *the series with positive terms* (although it would be more correct to call them the series with nonnegative terms). As to the series whose all terms are strictly greater than zero, we shall call them *the series with strictly positive terms.*

Series with positive terms are often encountered in applications. Moreover, their a priori knowledge will facilitate considerably the investigation of series with terms of any sign. In what follows, to emphasize the fact that we speak of a series with positive terms, we shall often designate the terms of such a series as p_k rather than u_k.

We can notice immediately the main specific feature of a series with positive terms: *the sequence of partial sums of such a series is nondecreasing.*

This enables us to make the following assertion.

Theorem 13.2. *For a series with positive terms to be convergent, it is necessary and sufficient for the sequence of its partial sums to be bounded.*

Necessity follows from the fact that every convergent sequence is bounded (by virtue of Theorem 3.8).

Sufficiency follows from the fact that the sequence of the partial sums does not decrease and, consequently, for this sequence to converge it must be bounded (by virtue of Theorem 3.15).

13.2.2. Comparison tests. In this subsection we shall establish a number of criteria allowing an inference to be made on a convergence (divergence) of the series being considered *by comparing it with another series whose convergence (or divergence) is a known fact.*

Theorem 13.3. *Assume that* $\sum\limits_{k=1}^{\infty} p_k$ *and* $\sum\limits_{k=1}^{\infty} p'_k$ *are two series with positive terms. Also suppose that for all the numbers k there holds an inequality*

$$p_k \leqslant p'_k. \tag{13.14}$$

Then the convergence of the series $\sum\limits_{k=1}^{\infty} p'_k$ *entails the convergence of the series* $\sum\limits_{k=1}^{\infty} p_k$, *the divergence of series* $\sum\limits_{k=1}^{\infty} p_k$ *entails the divergence of the series* $\sum\limits_{k=1}^{\infty} p'_k$.

Proof. Let us designate the nth partial sums of the series $\sum\limits_{k=1}^{\infty} p_k$ and $\sum\limits_{k=1}^{\infty} p'_k$ by S_n and S'_n respectively. From inequality (13.14) we infer that $S_n \leqslant S'_n$. The last inequality implies that the boundedness of the sequence of the partial sums $\{S'_n\}$ entails the boundedness of the sequence of the partial sums $\{S_n\}$ and, conversely, the unboundedness of the sequence of the partial sums $\{S_n\}$ entails the unboundedness of the sequence of the partial sums $\{S'_n\}$. By virtue of Theorem 13.2, Theorem 13.3 is proved.

Remark 1. We may put a condition into the hypothesis of Theorem 13.3 that inequality (13.14) should be satisfied not for all the numbers k but only *for those beginning with a certain number* k. In fact, in accordance with 13.1.3, the deletion of a finite number of terms does not affect the convergence of a series.

Remark 2. *Theorem* 13.3 *remains valid if in the hypothesis of this theorem we replace inequality* (13.14) *by the following inequality:*

$$p_k \leqslant cp'_k, \tag{13.15}$$

where c *is any positive constant.* Indeed, by virtue of 13.1.3, the question of the convergence of the series $\sum_{k=1}^{\infty} p'_k$ is equivalent to the question of the convergence of the series $\sum_{k=1}^{\infty} (cp'_k)$. In this case we can certainly put a requirement that inequality (13.15) must be complied with only beginning with a certain sufficiently large k.

Corollary of Theorem 13.3. If $\sum_{k=1}^{\infty} p_k$ *is a series with positive terms,* $\sum_{k=1}^{\infty} p'_k$ *is a series with strictly positive terms and if there exists a finite limit*

$$\lim_{k\to\infty} \frac{p_k}{p'_k} = L,$$

then the convergence of the series $\sum_{k=1}^{\infty} p'_k$ *leads to the convergence of the series* $\sum_{k=1}^{\infty} p_k$, *the divergence of the series* $\sum_{k=1}^{\infty} p_k$ *leads to the divergence of the series* $\sum_{k=1}^{\infty} p'_k$.

Proof. Since $\lim_{k\to\infty} \frac{p_k}{p'_k} = L$, from the definition of a limit it follows that for some $\varepsilon > 0$ there is a number N such that at $k \geqslant N$

$$L - \varepsilon < \frac{p_k}{p'_k} < L + \varepsilon.$$

Therefore, at $k \geqslant N$, there holds an inequality $p_k < (L + \varepsilon) p'_k$. The last inequality coincides with inequality (13.15) when $c = L + \varepsilon$. By virtue of Remark 2 to Theorem 13.3 the corollary is proved.

Theorem 13.4. Assume that $\sum_{k=1}^{\infty} p_k$ *and* $\sum_{k=1}^{\infty} p'_k$ *are two series with strictly positive terms. Assume also that for all numbers* k *there holds an*

inequality

$$\frac{p_{k+1}}{p_k} \leqslant \frac{p'_{k+1}}{p'_k}. \tag{13.16}$$

Then, the convergence of the series $\sum_{k=1}^{\infty} p'_k$ *entails the convergence of the series* $\sum_{k=1}^{\infty} p_k$ *and the divergence of the series* $\sum_{k=1}^{\infty} p_k$ *entails the divergence of the series* $\sum_{k=1}^{\infty} p'_k$.

Proof. Let us write inequality (13.16) for $k = 1, 2, \ldots, n-1$, where n is *any* number. We shall have

$$\frac{p_2}{p_1} \leqslant \frac{p'_2}{p'_1},$$
$$\frac{p_3}{p_2} \leqslant \frac{p'_3}{p'_2},$$
$$\ldots\ldots\ldots$$
$$\frac{p_n}{p_{n-1}} \leqslant \frac{p'_n}{p'_{n-1}}.$$

Multiplying term-by-term the inequalities written above, we obtain

$$\frac{p_n}{p_1} \leqslant \frac{p'_n}{p'_1} \text{ or } p_n \leqslant \frac{p_1}{p'_1} p'_n.$$

Since in the last inequality the quantity $c = p_1/p'_1$ is a *positive constant independent of the number* n, Theorem 13.4 is proved by virtue of Remark 2 to Theorem 13.3.

Remark 3. We may require in the hypothesis of Theorem 13.4 that inequality (13.16) should be satisfied not for all the numbers k but for those *beginning with a certain number* k (since the deletion of a finite number of the first terms does not affect the convergence of the series).

The two theorems proved above are known as the *comparison theorems* or the *comparison tests*.

Here are some examples of the application of the comparison tests.

1. Let us analyse the convergence of the series

$$\sum_{k=1}^{\infty} \frac{1}{3+b^k}, \text{ where } b > 0.$$

If $b \leqslant 1$, the kth term of the series being considered does not tend to zero as $k \to \infty$. Therefore, the necessary condition of the convergence of the series is violated and the series *diverges*. Now if $b > 1$,

then, the inequality

$$\frac{1}{3+b^k} < \frac{1}{b^k}$$

being valid for any number k and the series $\sum_{k=1}^{\infty} \frac{1}{b^k}$ being convergent, Comparison Theorem 13.3 enables us to assert that the series in question converges.

2. Let us decide whether the series

$$\sum_{k=1}^{\infty} \frac{1}{k^\alpha} = 1 + \frac{1}{2^\alpha} + \ldots + \frac{1}{k^\alpha} + \ldots \tag{13.17}$$

is convergent for any $\alpha \leqslant 1$. This series is often called a *generalized harmonic series.* Since the inequality

$$\frac{1}{k^\alpha} \geqslant \frac{1}{k}$$

is valid for any number k at $\alpha \leqslant 1$ and since the harmonic series $\sum_{k=1}^{\infty} \frac{1}{k}$ diverges *, Comparison Theorem 13.3 enables us to assert that series (13.17) diverges at *any* $\alpha \leqslant 1$.

13.2.3. D'Alembert's and Cauchy's tests. Two very applicable tests of convergence of series with positive terms, those by D'Alembert and by Cauchy, are closely related to the comparison tests. D'Alembert's and Cauchy's tests are based on the comparison of the series being considered with a series compiled of the elements of a geometric progression, namely, with the convergent series

$$\sum_{k=1}^{\infty} q^k = q + q^2 + q^3 + \ldots, \quad |q| < 1, \tag{13.18}$$

or with the divergent series

$$\sum_{k=1}^{\infty} 1 = 1 + 1 + 1 + \ldots \tag{13.19}$$

* We have established the divergence of a harmonic series in 13.1.2.

Theorem 13.5 (D'Alembert's* test). I. *If for all the numbers k or at least beginning with some number k, there holds an inequality*

$$\frac{p_{k+1}}{p_k} \leqslant q < 1^{**} \quad \left(\frac{p_{k+1}}{p_k} \geqslant 1\right), \tag{13.20}$$

then the series $\sum_{k=1}^{\infty} p_k$ *converges (diverges).*

II. *If there exists a limit*

$$\lim_{k\to\infty} \frac{p_{k+1}}{p_k} = L, \tag{13.21}$$

then the series $\sum_{k=1}^{\infty} p_k$ *converges at* $L < 1$ *and diverges at* $L > 1$. Theorem II is known as D'Alembert's test *in the limiting form.* In this form the test is most applicable.

Proof. Let us analyse separately Theorems I and II. (1) To prove Theorem I, let us put $p'_k = q^k$ $(p'_k = 1)$. Then we have $\frac{p'_{k+1}}{p'_k} = q$, where $q < 1$, $\left(\frac{p'_{k+1}}{p'_k} = 1\right)$, and we can rewrite inequality (13.20) in the form

$$\frac{p_{k+1}}{p_k} \leqslant \frac{p'_{k+1}}{p'_k} \quad \left(\frac{p_{k+1}}{p_k} \geqslant \frac{p'_{k+1}}{p'_k}\right). \tag{13.22}$$

Since the series $\sum_{k=1}^{\infty} p'_k$ coinciding with series (13.18) ((13.19)) converges (diverges), inequality (13.22), in accordance with Comparison Theorem 13.4, ensures the convergence (divergence) of the series $\sum_{k=1}^{\infty} p_k$. We have proved Theorem I.

(2) Let us now prove Theorem II. If $L < 1$, there is a *positive* number ε such that $L = 1 - 2\varepsilon$ and $L + \varepsilon = 1 - \varepsilon$. By the definition of the limit of a sequence, for the indicated ε there is a number N such that for $k \geqslant N$

$$L - \varepsilon < \frac{p_{k+1}}{p_k} < L + \varepsilon = 1 - \varepsilon. \tag{13.23}$$

The number $L + \varepsilon = 1 - \varepsilon$ plays the role of q in Theorem I. The series converges.

---

* Jean le Rond D'Alembert (1717-1783), a French mathematician and philosopher.

** In this case we assume, of course, that all the terms of the series $\sum_{k=1}^{\infty} p_k$ (at least beginning with a certain number) are *strictly* positive.

Now if $L > 1$, there is a *positive* number ε such that $L = 1 + \varepsilon$ and $L - \varepsilon = 1$. In this case, on the basis of the left-hand inequality (13.23), we obtain

$$\frac{p_{k+1}}{p_k} > L - \varepsilon = 1 \quad (\text{for } k \geqslant N).$$

The series diverges on the basis of Theorem I. The proof of Theorem 13.5 is complete.

Remarks to Theorem 13.5. (1) Let us pay attention to the fact that inequality $\frac{p_{k+1}}{p_k} \leqslant q < 1$ in Theorem 13.5 (I) (for all k beginning with a certain number) *cannot be replaced* by $\frac{p_{k+1}}{p_k} < 1$.

Indeed, as was proved above, the harmonic series (13.12) diverges, but for this series $\frac{p_{k+1}}{p_k} = \frac{k}{k+1} < 1$ (for all the numbers k).

(2) If in the conditions of Theorem 13.5 (II) $L = 1$, we cannot judge upon the convergence of the series (that is, D'Alembert's test does not work at $L = 1$). In fact, for the harmonic series (13.12) $L = 1$, and this series, as we know, is divergent. However, for the series

$$\sum_{k=1}^{\infty} \frac{1}{k^2} \tag{13.24}$$

we also have $L = 1$, but this series converges as will be shown in the next subsection.

Theorem 13.6 (Cauchy's test). I. *If the inequality*

$$\sqrt[k]{p_k} \leqslant q < 1 \left(\sqrt[k]{p_k} \geqslant 1\right) \tag{13.25}$$

holds for all the numbers k, or at least beginning with a certain number k, then the series $\sum_{k=1}^{\infty} p_k$ converges (diverges).

II. *If there exists a limit*

$$\lim_{k \to \infty} \sqrt[k]{p_k} = L, \tag{13.26}$$

then the series $\sum_{k=1}^{\infty} p_k$ converges for $L < 1$ and diverges for $L > 1$.

Theorem II is usually called Cauchy's test *in the limiting form.*

Proof. Let us analyse theorems I and II separately.

(1) To prove Theorem I, let us put $p'_k = q^k$ ($p'_k = 1$). Then, from inequality (13.25) we shall obtain

$$p_k \leqslant p'_k \ (p_k \geqslant p'_k). \tag{13.27}$$

Since the series $\sum_{k=1}^{\infty} p'_k$ coinciding with series (13.18) (13.19) converges (diverges), inequality (13.27) ensures the convergence (divergence) of the series $\sum_{k=1}^{\infty} p_k$ by virtue of Comparison Theorem 13.3.

We have proved Theorem 13.6 (I).

(2) To prove Theorem 13.6 (II), it is necessary to repeat verbatim the scheme we used to prove Theorem 13.5 (II), with the only exception of replacing $\frac{p_{k+1}}{p_k}$ by $\sqrt[k]{p_k}$ in the whole reasoning.

The proof of Theorem 13.6 is complete.

Remarks to Theorem 13.6. (1) As in the preceding theorem, the inequality $\sqrt[k]{p_k} \leqslant q < 1$ in Theorem 13.6 (I) cannot be replaced by $\sqrt[k]{p_k} < 1$.

(2) At $L = 1$, the Cauchy test in the limiting form "does not operate". Two examples can be cited which were indicated in the corresponding remark to D'Alembert's test.

(3) Which of the tests, then, D'Alembert's or Cauchy's, is stronger? Let us analyse this question as regards D'Alembert's and Cauchy's tests taken *in the limiting form*. It can be proved that *the existence of the limit* (13.21) *yields the existence of the limit* (13.26) *and the equality of these limits*. (The proof is presented in Supplement I to this chapter.) The converse is not true. Indeed, it is easy to verify that for the series

$$\sum_{k=1}^{\infty} \frac{(-1)^k + 3}{2^{k+1}} \tag{13.28}$$

the limit (13.26) exists and is equal to $\frac{1}{2}$ whereas the limit (13.21) does not exist at all. Thus we see that Cauchy's test is stronger than D'Alembert's test, for every time D'Alembert's test works, the test of Cauchy works too, while there are series (series (13.28), for instance) for which Cauchy's test works and D'Alembert's test does not. Notwithstanding all this, D'Alembert's test is used in practical applications more often than Cauchy's test.

Examples. (1) Let us investigate the problem of the convergence of the series

$$\sum_{k=1}^{\infty} \frac{(\sqrt{k})^k}{k!} . \tag{13.29}$$

Applying D'Alembert's test in the limiting form, we have

$$p_k = \frac{(\sqrt{k})^k}{k!}, \quad \frac{p_{k+1}}{p_k} = \frac{(\sqrt{k+1})^{k+1}}{(k+1)!} \cdot \frac{k!}{(\sqrt{k})^k} = \frac{1}{\sqrt{k+1}} \left(1 + \frac{1}{k}\right)^{k/2}. \tag{13.30}$$

On the basis of (13.30) we obtain

$$\lim_{k\to\infty}\frac{p_{k+1}}{p_k}=\lim_{k\to\infty}\frac{1}{\sqrt{k+1}}\left(1+\frac{1}{k}\right)^{k/2}=$$

$$=\lim_{k\to\infty}\frac{1}{\sqrt{k+1}}\lim_{k\to\infty}\left(1+\frac{1}{k}\right)^{k/2}=0\sqrt{e}=0<1,$$

that is, series (13.29) is convergent.

(2) Let us now investigate the problem of the convergence of the series

$$\sum_{k=1}^{\infty}\frac{k}{2^k}. \tag{13.31}$$

Applying Cauchy's test in the limiting form, we have

$$\sqrt[k]{p_k}=\frac{\sqrt[k]{k}}{2}. \tag{13.32}$$

By virtue of (13.32), $\lim\limits_{k\to\infty}\sqrt[k]{p_k}=\frac{1}{2}\lim\limits_{k\to\infty}\sqrt[k]{k}\,^{*}=\frac{1}{2}<1$. Therefore, Cauchy's test establishes the convergence of series (13.31).

13.2.4. Cauchy-Maclaurin's integral test. The tests of D'Alembert and Cauchy turn out to be inapplicable for elucidating the question concerning the convergence of some frequently encountered series with positive terms. Thus, for instance, these tests are helpless when we want to judge upon the convergence of the generalized harmonic series

$$\sum_{k=1}^{\infty}\frac{1}{k^{\alpha}} \tag{13.33}$$

(α is any real number).

Although we established at the end of 2.2 that at $\alpha\leqslant 1$ series (13.33) diverges, the problem remains unsolved as to the convergence of that series at $\alpha>1$. Here we shall establish one more general test for the convergence of a series with positive terms which will yield, in particular, the convergence of series (13.33) at $\alpha>1$.

Theorem 13.7 (Cauchy-Maclaurin's theorem). *Suppose the function $f(x)$ is nonnegative and does not increase everywhere on the half-line $x\geqslant m$, where m is any fixed number. Then the number series*

$$\sum_{k=m}^{\infty}f(k)=f(m)+f(m+1)+f(m+2)+\ldots \tag{13.34}$$

* To calculate $\lim\limits_{x\to+\infty}x^{1/x}$ it is necessary to take the logarithm of the expression $x^{1/x}$ and apply L'Hospital's rule.

converges if and only if there exists a limit, at $n \to \infty$, *of the sequence*

$$a_n = \int_m^n f(x)\,dx. \tag{13.35}$$

Proof. Assume that k is any number satisfying the condition $k \geqslant m+1$, and x is any value of the argument from the closed interval $k-1 \leqslant x \leqslant k$. Since by the hypothesis the function $f(x)$ does not increase on the indicated closed interval, the inequalities

$$f(k) \leqslant f(x) \leqslant f(k-1) \tag{13.36}$$

are valid for all x from the indicated interval. Being bounded and monotonic, the function $f(x)$ is integrable on the closed interval $k-1 \leqslant x \leqslant k$ (see 10.4.5). What is more, it follows from inequalities (13.36) and from property 3° (see 10.6.1) that

$$\int_{k-1}^{k} f(k)\,dx \leqslant \int_{k-1}^{k} f(x)\,dx \leqslant \int_{k-1}^{k} f(k-1)\,dx$$

or

$$f(k) \leqslant \int_{k-1}^{k} f(x)\,dx \leqslant f(k-1). \tag{13.37}$$

We have established inequalities (13.37) for any $k \geqslant m+1$. Let us now write these inequalities for the values of k equal to $m+1$, $m+2$, . . ., n, where n is any number exceeding m:

$$f(m+1) \leqslant \int_m^{m+1} f(x)\,dx \leqslant f(m),$$

$$f(m+2) \leqslant \int_{m+1}^{m+2} f(x)\,dx \leqslant f(m+1),$$

. . . . . .

$$f(n) \leqslant \int_{n-1}^{n} f(x)\,dx \leqslant f(n-1).$$

Adding up the written inequalities termwise, we obtain

$$\sum_{k=m+1}^{n} f(k) \leqslant \int_m^n f(x)\,dx \leqslant \sum_{k=m}^{n-1} f(k). \tag{13.38}$$

Let us agree to symbolize as S_n the nth sum of series (13.34) equal to

$$S_n = \sum_{k=m}^{n} f(k).$$

Having accepted this notation and taking into account the notation of (13.35), we can rewrite inequalities (13.38) as follows:

$$S_n - f(m) \leqslant a_n \leqslant S_{n-1}. \tag{13.39}$$

With the aid of inequalities (13.39) we can easily prove the theorem. In fact, it is evident from formula (13.35) that the sequence $\{a_n\}$ is nondecreasing. Consequently, for this sequence to converge, it is necessary and sufficient that it should be bounded. By virtue of Theorem 13.2, for series (13.34) to converge, it is necessary and sufficient that the sequence $\{S_n\}$ be bounded. It follows from inequalities (13.39) that the sequence $\{S_n\}$ is bounded if and only if the sequence $\{a_n\}$ is bounded, that is, if and only if the sequence $\{a_n\}$ converges. The proof of the theorem is complete.

Examples. (1) First of all, let us apply Cauchy-Maclaurin's integral test to investigate the convergence of the generalized harmonic series (13.33). Since (13.33) can be regarded as a series of the form of (13.34) at $m = 1$, $f(x) = 1/x^{\alpha}$ and the function $f(x)$ is decreasing and positive on the half-line $x \geqslant 1$, the question concerning the convergence of series (13.33) is equivalent to the question of the convergence of the sequence $\{a_n\}$, where

$$a_n = \int_1^n \frac{1}{x^{\alpha}}\,dx = \begin{cases} \dfrac{x^{1-\alpha}}{1-\alpha}\Big|_{x=1}^{x=n} = \dfrac{n^{1-\alpha}-1}{1-\alpha} & \text{for } \alpha \neq 1, \\ \ln x\big|_{x=1}^{x=n} = \ln n & \text{for } \alpha = 1. \end{cases}$$

It follows from the form of the terms a_n that the sequence $\{a_n\}$ diverges at $\alpha \leqslant 1$ and converges at $\alpha > 1$, with $\lim\limits_{n\to\infty} a_n = \dfrac{1}{\alpha-1}$ in the last case. Thus we see that *series* (13.33) *diverges at* $\alpha \leqslant 1$ (this was established earlier by another method) *and converges at* $\alpha > 1$. In particular, at $\alpha = 2$ series (13.33) passes into series (13.24) whose convergence we can now assert.

(2) Let us solve the problem of the convergence of the series

$$\sum_{k=2}^{\infty} \frac{1}{k \ln^{\beta} k}, \tag{13.40}$$

where β is a fixed positive real number. Series (13.40) can be considered as a series of the form of (13.34) at $m = 2$ and $f(x) = \dfrac{1}{x \ln^{\beta} x}$. Since the function $f(x)$ is nonnegative and does not

increase on the half-line $x \geqslant 2$, the problem of the convergence of series (13.40) is equivalent to that of convergence of the sequence $\{a_n\}$, where

$$a_n = \int_2^n \frac{1}{x \ln^\beta x}\, dx = \begin{cases} \dfrac{\ln^{1-\beta} x}{1-\beta}\Big|_{x=2}^{x=n} = \dfrac{\ln^{1-\beta} n - \ln^{1-\beta} 2}{1-\beta} & \text{at } \beta \neq 1, \\ \ln \ln x\Big|_{x=2}^{x=n} = \ln \ln n - \ln \ln 2 & \text{at } \beta = 1. \end{cases}$$

It follows from the form of the terms a_n that the sequence $\{a_n\}$ converges at $\beta > 1$ and diverges at $\beta \leqslant 1$. Thus we see that series (13.40) *converges at* $\beta > 1$ *and diverges at* $\beta \leqslant 1$.

13.2.5. Raabe's test. It should be recalled that D'Alembert's and Cauchy's tests were based on the comparison between the series under consideration and a series constituting the sum of a geometric progression. It is naturally desirable to obtain some more refined criteria based on the comparison between the series being considered and other standard series converging or diverging "at a slower rate" than the series for a geometric progression.

We shall establish here a test based on the comparison between the series being considered and the standard series

$$\sum_{k=1}^{\infty} \frac{1}{k^\alpha} = 1 + \frac{1}{2^\alpha} + \frac{1}{3^\alpha} + \dots \tag{13.41}$$

we have studied above.

Theorem 13.8 (Raabe's* test). I. *If for all the numbers* k, *or at least from a certain number* k, *the inequality*

$$k\left(1 - \frac{p_{k+1}}{p_k}\right) \geqslant q > 1^{**} \quad \left\{k\left(1 - \frac{p_{k+1}}{p_k}\right) \leqslant 1\right\}, \tag{13.42}$$

holds true, then the series $\sum_{k=1}^{\infty} p_k$ *converges (diverges).*

II. *If there exists a limit*

$$\lim_{k \to \infty} k\left(1 - \frac{p_{k+1}}{p_k}\right) = L, \tag{13.43}$$

then the series $\sum_{k=1}^{\infty} p_k$ *converges at* $L > 1$ *and diverges at* $L < 1$. Theorem II is known as Raabe's test *in the limiting form.*

Proof. Let us consider separately Theorems I and II.

(1) To prove Theorem I, we shall rewrite inequality (13.42) in the form

$$\frac{p_{k+1}}{p_k} \leqslant 1 - \frac{q}{k} \quad \left\{\frac{p_{k+1}}{p_k} \geqslant 1 - \frac{1}{k}\right\}. \tag{13.44}$$

Since $q > 1$, there is a number α satisfying the inequalities $q > \alpha > 1$. Expanding the function $(1 + x)^\alpha$ by Maclaurin's formula with the remainder in

* Josef Ludwig Raabe (1801-1859), a Swiss mathematician.

** In this case it is certainly assumed that the series $\sum_{k=1}^{\infty} p_k$ has *strictly positive terms*, at least beginning with a certain number.

Peano's form (see 8.15.2), we shall have

$$(1+x)^{\alpha}=1+\alpha x+o(x).$$

Putting $x=-\frac{1}{k}$ in the last formula, we obtain

$$\left(1-\frac{1}{k}\right)^{\alpha}=1-\frac{\alpha}{k}+o\left(\frac{1}{k}\right). \tag{13.45}$$

Since the sequence $\dfrac{o\left(\frac{1}{k}\right)}{\frac{1}{k}}$ is infinitesimal, the inequality

$$\frac{o\left(\frac{1}{k}\right)}{\frac{1}{k}}\leqslant q-\alpha \tag{13.46}$$

is valid beginning with a certain number k_0. Comparing (13.45) and (13.46), we obtain an inequality

$$\left(1-\frac{1}{k}\right)^{\alpha}\geqslant 1-\frac{q}{k} \quad (\text{for } k\geqslant k_0). \tag{13.47}$$

A comparison of inequalities (13.44) and (13.47) yields

$$\frac{p_{k+1}}{p_k}\leqslant\left(1-\frac{1}{k}\right)^{\alpha}\left\{\frac{p_{k+1}}{p_k}\geqslant 1-\frac{1}{k}\right\} \quad (\text{for } k\geqslant k_0).$$

The last inequalities can be rewritten in the form

$$\frac{p_{k+1}}{p_k}\leqslant\frac{\frac{1}{k^{\alpha}}}{\frac{1}{(k-1)^{\alpha}}}\left\{\frac{p_{k+1}}{p_k}\geqslant\frac{\frac{1}{k}}{\frac{1}{k-1}}\right\} \quad (\text{for } k\geqslant k_0). \tag{13.48}$$

Since series (13.41) converges at $\alpha>1$ and diverges at $\alpha=1$, inequalities (13.48) and Comparison Theorem 13.4 enable us to assert that the series $\sum\limits_{k=1}^{\infty}p_k$ converges (diverges). Thus we have proved Theorem I.

(2) In the same way as in D'Alembert's and Cauchy's tests we shall reduce Theorem II to Theorem I. First assume that $L>1$. Let us put $\varepsilon=\dfrac{L-1}{2}$, $q=1+\varepsilon=L-\varepsilon$. By the definition of the limit (13.43), we can indicate for this number ε a number k_0 such that beginning with this k_0 we have $\left|k\left(1-\frac{p_{k+1}}{p_k}\right)-L\right|<\varepsilon$ and, consequently, the left-hand inequality (13.42) holds true. If now $L<1$, we shall put $\varepsilon=1-L$ and, making use of the definition of the limit (13.43), find that beginning with some number k_0 the right-hand inequality (13.42) holds true. The proof of Theorem 13.8 is complete.

Remark. Note that in Theorem 13.8 (I) we cannot take $q=1$ in the left-hand inequality (13.42) (in this case the series may not be convergent). At $L=1$

Theorem 13.8 (II) does not work (both convergence and divergence are possible).

Example. It is required to investigate the question as to the convergence of the series

$$\sum_{k=2}^{\infty} p_k, \text{ where } p_k = a^{-\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{k-1}\right)} \quad (a=\text{const}>0).$$

It is easy to verify that when applied to this series D'Alembert's and Cauchy's tests do not work. Let us apply Raabe's test. It is easy to verify that

$$k\left(1-\frac{p_{k+1}}{p_k}\right)=\frac{a^{-\frac{1}{k}}-1}{\left(-\frac{1}{k}\right)}.$$

It is not difficult to realize that at $k\to\infty$ the last fraction tends to the derivative of the function a^x at point $x=0$, that is, to $\ln a$. By virtue of Raabe's test, the series in question *converges at* $\ln a>1$, *i.e. at* $a>e$, *and diverges at* $\ln a<1$, *i.e. at* $a<e$. When $a=e$, the question as to the convergence of the series requires additional investigation, since Raabe's test does not work. Series (13.40) can serve as another example of a series to which Raabe's test is not applicable.

13.2.6. Absence of a universal series of comparison. As was pointed out earlier, D'Alembert's and Cauchy's tests are based on the comparison between the series being considered and a series for a geometric progression, while Raabe's test is based on the comparison with a more slowly converging (or diverging) series (13.41).

A question naturally arises *whether there exists a universal (exceptionally slowly!) converging (or diverging) series a comparison with which would make it possible to make an inference on the convergence (or divergence) of any preassigned series with positive terms.*

We shall prove that *there is no* universal series of this kind. Suppose we are given two convergent series $\sum_{k=1}^{\infty} p_k$ and $\sum_{k=1}^{\infty} p'_k$. Their respective nth remainders will be designated as r_n and r'_n. We shall say that *the series* $\sum_{k=1}^{\infty} p'_k$ *converges slower than the series* $\sum_{k=1}^{\infty} p_k$ *if* $\lim_{n\to\infty}\frac{r_n}{r'_n}=0$. We shall prove that *for every convergent series there exists a series converging slower than the indicated series.* Indeed, assume that $\sum_{k=1}^{\infty} p_k$ is any converging series and r_n is its nth remainder. Let us prove that the series $\sum_{k=1}^{\infty} p'_k$, where * $p'_k=$ $=\sqrt{r_{k-1}}-\sqrt{r_k}$, converges slower than the series $\sum_{k=1}^{\infty} p_k$. In fact, if r'_n is

* The whole sum $\sum_{k=1}^{\infty} p_k$ is taken as r_0.

the nth remainder of the series $\sum_{k=1}^{\infty} p'_k$, then

$$\lim_{n\to\infty} \frac{r_n}{r'_n} = \lim_{n\to\infty} \frac{r_n}{\sqrt{r_n}} = 0.$$

Let us prove now the absence of a universal convergent series a comparison with which would make it possible to make a conclusion on the convergence of any preassigned convergent series. Indeed, if there existed such a universal convergent series $\sum_{k=1}^{\infty} p_k$, then having taken for it the series $\sum_{k=1}^{\infty} p'_k$ constructed above, we would have

$$\lim_{k\to\infty} \frac{p_k}{p'_k} = \lim_{k\to\infty} \frac{r_{k-1} - r_k}{\sqrt{r_{k-1}} - \sqrt{r_k}} = \lim_{k\to\infty} (\sqrt{r_{k-1}} + \sqrt{r_k}) = 0.$$

Thus *we cannot judge upon the convergence of the series* $\sum_{k=1}^{\infty} p'_k$ *by comparing it to the series* $\sum_{k=1}^{\infty} p_k$. By a complete analogy we can prove the absence of a universal divergent series a comparison with which would allow us to make a conclusion about the divergence of any preassigned divergent series.

13.3. ABSOLUTELY AND CONDITIONALLY CONVERGENT SERIES

13.3.1. The concept of an absolutely and conditionally convergent series. We shall now investigate series whose terms are real numbers of any sign.

Definition 1. *We shall call the series*

$$\sum_{k=1}^{\infty} u_k \tag{13.49}$$

absolutely convergent *if the series*

$$\sum_{k=1}^{\infty} |u_k| \tag{13.50}$$

converges. Note that this definition does not contain a supposition concerning the convergence of series (13.49) itself. Such a supposition turns out to be superfluous since the following theorem is valid.

Theorem 13.9. *The convergence of series* (13.50) *implies the convergence of series* (13.49).

Proof. Let us make use of Cauchy's test for a series (i.e. Theorem 13.1). We have to prove that for any $\varepsilon > 0$ there is a number N such that for all the numbers n complying with the condition $n \geqslant N$

and for any natural p

$$\left|\sum_{k=n+1}^{n+p} u_k\right| < \varepsilon. \tag{13.51}$$

We take any $\varepsilon > 0$. Since series (13.50) is convergent, there is a number N, by virtue of Theorem 13.1, such that for all the numbers n satisfying the condition $n \geqslant N$ and for any natural p

$$\sum_{k=n+1}^{n+p} |u_k| < \varepsilon. \tag{13.52}$$

Bearing in mind that the absolute value of the sum of several summands does not exceed the sum of their absolute values, we can write

$$\left|\sum_{k=n+1}^{n+p} u_k\right| \leqslant \sum_{k=n+1}^{n+p} |u_k|. \tag{13.53}$$

Comparing inequalities (13.52) and (13.53) we obtain inequality (13.51). The theorem has been proved.

Definition 2. *Series* (13.49) *is called* **conditionally convergent** *if that series converges whereas the corresponding series of the absolute values* (13.50) *diverges.*

We can cite the following series as an example of an *absolutely* convergent series:

$$\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k^{\alpha}} = 1 - \frac{1}{2^{\alpha}} + \frac{1}{3^{\alpha}} - \frac{1}{4^{\alpha}} + \ldots, \quad \text{where} \quad \alpha > 1.$$

This series converges absolutely since series (13.33) converges for $\alpha > 1$. An example is due of a *conditionally* convergent series. We shall prove the conditional convergence of the series

$$\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots + \frac{(-1)^{n-1}}{n} + \ldots \tag{13.54}$$

Since, as we know, the corresponding series of absolute values (harmonic series) *diverges*, it follows that to prove the conditional convergence of series (13.54), it is sufficient to show that this series converges. Let us prove that series (13.54) converges to the number ln 2. In 8.15.2 we have obtained the expansion of the function ln $(1 + x)$ by Maclaurin's formula:

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots + (-1)^{n-1}\frac{x^n}{n} + R_{n+1}(x). \tag{13.55}$$

In the same section we have obtained the following estimation of the remainder for all x from the closed interval $0 \leqslant x \leqslant 1$:

$$|R_{n+1}(x)| < \frac{1}{n+1}.$$

Putting $x = 1$ in formulas (13.55) and (13.56), we have

$$\ln 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{(-1)^{n-1}}{n} + R_{n+1}(1),$$

where

$$|R_{n+1}(1)| < \frac{1}{n+1}$$

or

$$\left|\left[1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots + \frac{(-1)^{n-1}}{n}\right] - \ln 2\right| < \frac{1}{n+1}.$$

Denoting the nth partial sum of series (13.54) by S_n, we can rewrite the last inequality as

$$|S_n - \ln 2| < \frac{1}{n+1}.$$

Thus the difference $S_n - \ln 2$ is an infinitesimal sequence and this proves the convergence of series (13.54) to the number ln 2.

13.3.2. On rearrangement of terms of a conditionally convergent series. One of the most important properties of the sum of a finite number of real summands is the *commutativity property* which asserts that a rearrangement of the summands does not affect the sum. A question naturally arises whether this property remains valid for the sum of a convergent series, that is, *whether the sum of a convergent series can be affected by a rearrangement of the terms of the series.* Here we shall elucidate this question with regard to a *conditionally convergent series.* We shall begin with the investigation of a specific rearrangement of the terms of series (13.54). For the sake of convenience, we shall write series (13.54) in the form

$$\underbrace{1 - \frac{1}{2}}_{} + \underbrace{\frac{1}{3} - \frac{1}{4}}_{} + \ldots + \underbrace{\frac{1}{2k-1} - \frac{1}{2k}}_{} + \ldots \qquad (13.56)$$

We have proved at the end of the preceding subsection that series (13.54) is conditionally convergent and possesses a sum $S = \ln 2$. Let us now rearrange the terms of series (13.54) to have two negative terms after one positive term. As a result we shall have

$$\underbrace{1 - \frac{1}{2} - \frac{1}{4}}_{} + \underbrace{\frac{1}{3} - \frac{1}{6} - \frac{1}{8}}_{} + \ldots + \left(\frac{1}{2k-1} - \frac{1}{4k-2} - \frac{1}{4k}\right) + \ldots \qquad (13.57)$$

We shall prove that series (13.57) obtained as a result of the rearrangement of the terms of series (13.54) converges and possesses a sum half that of series (13.54). We shall use the symbols S_m and S'_m, respectively, for the mth partial sums of series (13.54) and (13.57). We can then write

$$S'_{3m}=\sum_{k=1}^{m}\left(\frac{1}{2k-1}-\frac{1}{4k-2}-\frac{1}{4k}\right)=\sum_{k=1}^{m}\left(\frac{1}{4k-2}-\frac{1}{4k}\right)=$$

$$=\frac{1}{2}\sum_{k=1}^{m}\left(\frac{1}{2k-1}-\frac{1}{2k}\right)=\frac{1}{2}S_{2m}.$$

Thus we have

$$S'_{3m}=\frac{1}{2}S_{2m}. \tag{13.58}$$

And then it is evident that

$$S'_{3m-1}=\frac{1}{2}S_{2m}+\frac{1}{4m}, \tag{13.59}$$

$$S'_{3m-2}=S'_{3m-1}+\frac{1}{4m-2}. \tag{13.60}$$

Since $\lim\limits_{m\to\infty} S_{2m}=S$, we shall obtain, in the limit, as $m\to\infty$, from formulas (13.58), (13.59) and (13.60)

$$\lim_{m\to\infty} S'_{3m}=\frac{1}{2}S,\ \lim_{m\to\infty} S'_{3m-1}=\frac{1}{2}S,\ \lim_{m\to\infty} S'_{3m-2}=\frac{1}{2}S.$$

Thus we have obtained a complete proof of the fact that series (13.57) converges and has a sum equal to (1/2) S. Since $S=\ln 2\neq 0$, it is evident that (1/2) $S\neq S$. Hence, *as a result of the rearrangement of the terms indicated above the sum of the conditionally convergent series* (13.54) *has changed.* The concrete example we have considered shows that a conditionally convergent series *does not possess the commutativity property.* The following remarkable assertion belonging to Riemann fully clarifies the problem concerning the influence of the rearrangement of terms on the sum of a conditionally convergent series.

Theorem 13.10 (Riemann's theorem). *If a series is conditionally convergent, then, whatever the preassigned number L, the terms of the series can be rearranged so that the rearranged series will converge to the number L.*

Proof. Assume that

$$\sum_{k=1}^{\infty} u_k \tag{13.61}$$

is an arbitrary conditionally convergent series. Let us denote by $p_1, p_2, p_3, \ldots$ the *positive terms* of series (13.61) written in the order of their appearance in the series, and by $q_1, q_2, q_3, \ldots$ the *absolute values of the negative terms* of series (13.61) written in the order in which they appear in that series. Series (13.61) contains an *infinitude of positive and negative terms* for if there were a finite number of terms of one definite sign, then, by deleting a finite number of the first terms not affecting the convergence, we would obtain a series consisting of terms of like signs and its convergence would be *absolute*. Thus, connected with series (13.61) are two infinite series with *positive* terms $\sum_{k=1}^{\infty} p_k$ and $\sum_{k=1}^{\infty} q_k$. We shall use the symbol P to denote the first of these series and the symbol Q for the second series. Let us prove that both series, P and Q, are divergent. We shall designate by S_n the nth partial sum of series (13.61), by P_n, the sum of all positive terms of S_n, and by Q_n, the sum of the absolute values of all the negative terms of S_n. Then, evidently, $S_n = P_n - Q_n$, and since by the hypothesis series (13.61) converges to some number S, we have

$$\lim_{n\to\infty} (P_n - Q_n) = S. \tag{13.62}$$

On the other hand, since series (13.61) *is not absolutely convergent*, it follows that

$$\lim_{n\to\infty} (P_n + Q_n) = \infty. \tag{13.63}$$

Comparing (13.62) and (13.63), we obtain $\lim_{n\to\infty} P_n = \infty$, $\lim_{n\to\infty} Q_n = \infty$, that is, we have proved that both series P and Q are divergent. From their divergence it follows that even after the deletion of a finite number of the first terms of these series, we can take from the remaining terms of series P, as well as series Q, as many terms as to make their sum exceed any preassigned number. Proceeding from this fact we shall prove that it is possible to rearrange the terms of the original series (13.61) so that the resulting series will converge to the preassigned number L. Indeed, we shall obtain the required series in the following way. First we shall choose from the original series (13.61) *as many* positive terms $p_1, p_2, p_3, \ldots, p_{k_1}$ *as is required* for the sum $p_1 + p_2 + \ldots + p_{k_1}$ to exceed L. Then we shall add to the chosen terms *as many* negative terms $-q_1, -q_2, \ldots, -q_{k_2}$ as to make the total sum $p_1 + p_2 + \ldots + p_{k_1} - q_1 - q_2 - \ldots - q_{k_2}$ smaller than L. Then again we shall add *as many* positive terms $p_{k_1+1}, p_{k_1+2}, \ldots, p_{k_2}$ as to make the total sum $p_1 + p_2 + \ldots + p_{k_1} - q_1 - q_2 - \ldots - q_{k_2} + p_{k_1+1} + \ldots + p_{k_3}$ larger than L. Continuing the argument we shall obtain an infinite series containing *all the terms* of the original series

(13.61) since each time we would have to add *at least one* positive or negative term of the original series. It remains to prove that the resulting series converges to L. It should be noted that in the series we have obtained *groups of positive terms* alternate with *groups of negative terms*. If the partial sum of the series obtained ends with a *complete group*, then the deviation of this partial sum from the number L does not exceed the absolute value of its last term*. Now if the partial sum ends with a *noncomplete group*, the deviation of this partial sum from the number L does not exceed the absolute value of the last term of the last but one group. To verify the convergence of the series to L, it is sufficient to make sure that the absolute values of the last terms of the groups form an infinitesimal sequence, and this follows immediately from the necessary condition of convergence of the original series (13.61). The proof of Riemann's theorem is complete.

13.3.3. On rearrangement of an absolutely convergent series. We have proved in the preceding subsection that *a conditionally convergent series does not possess a commutation property*. Here we shall prove that the *commutativity property is valid for every absolutely convergent series.*

Theorem 13.11 (Cauchy's theorem). *If a given series is absolutely convergent, then any series obtained from the given series by means of some rearrangement of the terms is also absolutely convergent and possesses the same sum as the given series.*

Proof. Assume that the series

$$\sum_{k=1}^{\infty} u_k \tag{13.64}$$

is absolutely convergent and the sum of this series is equal to S. Assume, in addition, that the series

$$\sum_{k=1}^{\infty} u'_k \tag{13.65}$$

is a series obtained from series (13.64) by some rearrangement of the terms. It is required to prove that (1) series (13.65) converges and has a sum equal to S, (2) series (13.65) is absolutely convergent. We shall first prove requirement (1). It suffices to prove that for any $\varepsilon > 0$ there is a number N such that for $n \geqslant N$

$$\left| \sum_{k=1}^{n} u'_k - S \right| < \varepsilon. \tag{13.66}$$

We take an arbitrary $\varepsilon > 0$. Since series (13.64) is absolutely convergent and possesses a sum equal to S, there can be indicated a

* Since we add terms to the given group until the total sum "goes over" the number L.

number N_0 for the chosen $\varepsilon > 0$, such that there will hold inequalities

$$\sum_{k=N_0+1}^{N_0+p} |u_k| < \frac{\varepsilon}{2} \text{ (p is any natural number)} \tag{13.67}$$

and

$$\left| \sum_{k=1}^{N_0} u_k - S \right| < \frac{\varepsilon}{2}^{*}. \tag{13.68}$$

Let us now choose a number N large enough for any partial sum S_n of series (13.65) with a number n exceeding N *to contain all the first* N_0 *terms of series* (13.65)**.

Let us evaluate the difference on the left-hand side of (13.66) and prove that for $n \geqslant N$ inequality (13.66) is valid for this difference.

Actually, the indicated difference can be represented as

$$\sum_{k=1}^{n} u'_k - S = \left(\sum_{k=1}^{n} u'_k - \sum_{k=1}^{N_0} u_k\right) + \left(\sum_{k=1}^{N_0} u_k - S\right). \tag{13.69}$$

Since the absolute value of the sum of two quantities does not exceed the sum of their absolute values, we obtain from (13.69)

$$\left| \sum_{k=1}^{n} u'_k - S \right| \leqslant \left| \sum_{k=1}^{n} u'_k - \sum_{k=1}^{N_0} u_k \right| + \left| \sum_{k=1}^{N_0} u_k - S \right|. \tag{13.70}$$

It is evident from inequalities (13.68) and (13.70) that to prove inequality (13.66) it is sufficient to prove that for $n \geqslant N$

$$\left| \sum_{k=1}^{n} u'_k - \sum_{k=1}^{N_0} u_k \right| < \frac{\varepsilon}{2}. \tag{13.71}$$

To prove inequality (13.71), let us note that for $n \geqslant N$ the first of the sums on the left-hand side of (13.71) *contains all* N_0 *first terms of series* (13.64). Consequently, the difference

$$\sum_{k=1}^{n} u'_k - \sum_{k=1}^{N_0} u_k \tag{13.72}$$

is the sum of $(n - N_0)$ terms of series (13.64) with the numbers *each of which exceeds* N_0.

---

* The number N_0 in inequalities (13.67) and (13.68) can be *the same*. Indeed, having written the two indicated inequalities with different numbers N_0, we can then choose the larger of the two numbers N_0.

** Such a number N can be chosen since series (13.65) is obtained from series (13.64) by means of some rearrangement of the terms.

If we choose the natural p large enough for the number $N_0 + p$ *to exceed the numbers of all* $(n - N_0)$ *terms of the sum just mentioned*, then for the difference (13.72) there obviously holds an inequality

$$\left| \sum_{k=1}^{n} u'_k - \sum_{k=1}^{N_0} u_k \right| \leqslant \sum_{k=N_0+1}^{N_0+p} |u_k|. \tag{13.73}$$

Inequalities (13.73) and (13.67) yield inequality (13.71). Thus we have proved inequality (13.66), that is showed that series (13.65) converges and possesses a sum equal to S. It remains to prove assertion (2) stating that series (13.65) is *absolutely* convergent. The proof of this assertion will follow from assertion (1) if we apply it to the series

$$\sum_{k=1}^{\infty} |u_k| \text{ and } \sum_{k=1}^{\infty} |u'_k|. \tag{13.74}$$

While doing this, we shall prove the convergence of the second series (13.74), i.e. the absolute convergence of series (13.65). The proof of Theorem 13.11 is complete.

13.4. ARITHMETIC OPERATIONS ON CONVERGENT SERIES

We shall discuss here the problem of a possible termwise addition and multiplication of convergent series.

Theorem 13.12. *If two series* $\sum_{k=1}^{\infty} u_k$ *and* $\sum_{k=1}^{\infty} v_k$ *converge and possess sums equal, respectively, to* U *and* V, *then the series* $\sum_{k=1}^{\infty} (u_k \pm v_k)$ *converges and possesses a sum equal to* $U \pm V$.

Proof. Let us designate the nth partial sums of the series $\sum u_k$, $\sum v_k$ and $\sum (u_k \pm v_k)$ by U_n, V_n and S_n respectively. Then, evidently, $S_n = U_n \pm V_n$. Since $\lim_{n\to\infty} U_n = U$, $\lim_{n\to\infty} V_n = V$, it follows that in accordance with Theorems 3.9 and 3.10 there exists a limit $\lim_{n\to\infty} S_n = U \pm V$. The theorem has been proved.

Thus we see that *any convergent series can be added and subtracted term by term.*

Before discussing the problem of a termwise multiplication of series, we shall prove the following assertion.

Theorem 13.13. *If two series* $\sum_{k=1}^{\infty} u_k$ *and* $\sum_{l=1}^{\infty} v_l$ *are absolutely convergent and possess sums equal, respectively, to* U *and* V, *then the series compiled from all the products of the form* $u_k v_l$ $(k = 1, 2, \ldots;$ $l = 1, 2, \ldots)$, *enumerated in an arbitrary order, is also absolutely convergent and its sum is equal to* UV.

Proof. Designating by $w_1, w_2, w_3, \ldots$ the products of the form $u_k v_l$ $(k = 1, 2, \ldots; l = 1, 2, \ldots)$, enumerated in an arbitrary order, we shall prove that the series $\sum_{i=1}^{\infty} |w_i|$ is convergent. Assume that S_n is the nth partial sum of the series. The sum S_n consists of the terms of the form $|u_k v_l|$. Among the indices k and l of such terms entering into the sum S_n there can be found the *greatest* index which will be denoted by m. Then, in any case,

$$S_n \leqslant (|u_1| + |u_2| + \ldots + |u_m|)(|v_1| + |v_2| + \ldots + |v_m|). \tag{13.75}$$

The right-hand side of inequality (13.75) contains the product of the mth partial sums of the series $\sum |u_k|$ and $\sum |v_l|$. By virtue of the convergence of the indicated series with positive terms, all their partial sums (and, consequently, their products) are *bounded*. Therefore, the sequence of the partial sums $\{S_n\}$ is *bounded*, and this proves the convergence of the series $\sum |w_i|$, i.e. the *absolute* convergence of the series $\sum w_i$.

It remains to prove that the last series possesses a sum S equal to UV. Since this series is absolutely convergent, it follows by virtue of Theorem 13.11 that its sum S *does not depend on the order of its summation*. Whatever the sequence (and, consequently, the *subsequence**) of the partial sums of that series, it converges to the number S. But in that case the sum S of the series $\sum_{i=1}^{\infty} w_i$ is obviously equal to UV for it is precisely to this number that the *subsequence* W_m of the partial sums of this series converges, the *subsequence* W_m being of the form

$$W_m = (u_1 + u_2 + \ldots + u_m)(v_1 + v_2 + \ldots + v_m).$$

We have proved Theorem 13.13.

Remark. For various purposes it is more convenient to write the product of the series $\sum_{k=1}^{\infty} u_k$ and $\sum_{k=1}^{\infty} v_k$ in the form

$$\left(\sum_{k=1}^{+\infty} u_k\right)\left(\sum_{k=1}^{+\infty} v_k\right) = u_1 v_1 + (u_1 v_2 + u_2 v_1) + \ldots$$
$$\ldots + (u_1 v_{k-1} + \ldots + u_{k-1} v_1) + \ldots$$

We wish to point out the fact, without proving it, that the series obtained by termwise multiplication of two series using a special method indicated above is also convergent in the case when *only one* of the two series being multiplied is *absolutely* convergent (the

* By virtue of 3.4.1.

other series may be only conditionally convergent). In the case where both series are conditionally convergent, their termwise multiplication even by this method leads, in general, to a divergent series.

13.5. TESTS OF CONVERGENCE OF ARBITRARY SERIES

We have established in Sec. 2 a number of tests for convergence for series *with positive terms*. Here we shall study the tests for convergence for series with terms of any sign. Thus, suppose that

$$\sum_{k=1}^{\infty} u_k \tag{13.76}$$

is a series whose terms are of arbitrary signs. Note first of all that to establish the *absolute* convergence of that series, that is, to establish the convergence of a series with positive terms

$$\sum_{k=1}^{\infty} |u_k|,$$

use can be made of any of the tests presented in 13.2 (the tests of D'Alembert, Cauchy, Raabe or the integral test). But none of the indicated tests makes it possible to elucidate the *more refined question concerning the conditional convergence of series* (13.76)*.

We shall now try to find more refined test making it possible to establish the convergence of series (13.76) also in the cases where the series is not absolutely convergent.

13.5.1. Leibniz's test. Leibniz's test is a particular form of series (13.76), most often applied in practice, the so-called *alternating*

* It should be pointed out, however, that D'Alembert's and Cauchy's tests can be used *to establish the divergence of aseries with terms of any sign* (13.76). Indeed, every time D'Alembert's or Cauchy's test establishes the divergence of the series of the absolute values $\sum_{k=1}^{\infty} |u_k|$, the kth term u_k of series (13.76) does not tend to zero as $k \to \infty$, i.e. series (13.76) diverges. We shall verify, as an example, that the series $\sum_{k=1}^{\infty} k! \left(\frac{x}{k}\right)^k$ diverges for any fixed value of x satisfying the inequality $|x| > e$. It should be emphasized that it is difficult to verify directly that the kth term of the series in question does not tend to zero as $k \to \infty$. Let us apply D'Alembert's test to the series being considered. Denoting the kth term of the series by a_k, we shall have $\frac{|a_{k+1}|}{|a_k|} = \frac{|x|}{\left(1+\frac{1}{k}\right)^k}$, whence we have $\lim_{k\to\infty} \frac{|a_{k+1}|}{|a_k|} = \frac{|x|}{e} > 1$. We have proved the divergence of the series.

series. A series is called *alternating* if its terms are alternately positive and negative. It is convenient to write out an alternating series in the form making the signs of all its terms evident, i.e. in the form

$$p_1 - p_2 + p_3 - \ldots + (-1)^{k-1} p_k + \ldots, \tag{13.77}$$

where all $p_k \geqslant 0$.

Theorem 13.14 (Leibniz's test). *If the terms of an alternating series, being taken in absolute values, form a nonincreasing infinitesimal sequence, then the series is convergent.*

Remark 1. A series satisfying the conditions of Theorem 13.14 is often called *Leibniz's series.*

Proof of Theorem 13.14. Suppose we have a series (13.77) and it is known that the sequence $\{p_k\}$ is nonincreasing and infinitesimal. A partial sum of this series of an *even* order S_{2n} can be written as

$$S_{2n} = (p_1 - p_2) + (p_3 - p_4) + \ldots + (p_{2n-1} - p_{2n}). \tag{13.78}$$

Since every expression in parentheses in (13.78) is *nonnegative**, it is clear that with an increase in n the sequence $\{S_{2n}\}$ *does not decrease.*

On the other hand, S_{2n} can be rewritten in the form

$$S_{2n} = p_1 - (p_2 - p_3) - (p_4 - p_5) - \ldots$$
$$\ldots - (p_{2n-2} - p_{2n-1}) - p_{2n},$$

whence it is evident that $S_{2n} \leqslant p_1$ for any number n. Thus we see that the sequence of *even* partial sums S_{2n} does not decrease and is bounded above. By virtue of Theorem 3.15 this sequence converges to some number S, i.e. $\lim\limits_{n\to\infty} S_{2n} = S$. It follows from the obvious equality $S_{2n-1} = S_{2n} + p_{2n}$ and from the fact that $\lim\limits_{n\to\infty} p_{2n} = 0$ that the sequence of *odd* partial sums $\{S_{2n-1}\}$, too, converges to the same number S, i.e. $\lim\limits_{n\to\infty} S_{2n-1} = S$. Thus, the whole sequence $\{S_n\}$ converges to S. The theorem has been proved.

Remark 2. In the process of proving Theorem 13.14 we have revealed that the sequence of *even* partial sums $\{S_{2n}\}$ converges to the limit S *without decreasing.* Similarly, it follows from the equality

$$S_{2n-1} = p_1 - (p_2 - p_3) - (p_4 - p_5) - \ldots$$
$$\ldots - (p_{2n-2} - p_{2n-1})$$

that the sequence of odd partial sums $\{S_{2n-1}\}$ converges to the limit S *without increasing.*

---

* Because $\{p_k\}$ does not increase, i.e. $p_k \geqslant p_{k+1}$.

Thus, for any number n

$$S_{2n} \leqslant S \leqslant S_{2n-1}. \tag{13.79}$$

Since $S_{2n-1} - S_{2n} = p_{2n}$, it follows from inequalities (13.79) that

$$S - S_{2n} \leqslant p_{2n}$$

and

$$S_{2n-1} - S \leqslant p_{2n} \leqslant p_{2n-1}.$$

Thus we have found that for any number n there holds an inequality

$$|S_n - S| \leqslant p_n \tag{13.80}$$

which is often employed for approximate calculations with the aid of series.

By way of examples, let us consider the series

$$\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots + \frac{(-1)^{k-1}}{k} + \ldots \tag{13.81}$$

which was discussed above more than once. Note that series (13.81) is Leibniz's series and, therefore, its convergence follows from Theorem 13.14. Suppose, for instance, we have to calculate the sum of series (13.81), i.e. the number ln 2, with an accuracy to within $\frac{1}{10^n}$. By virtue of estimation (13.80), this sum coincides with $S_{10^n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots - \frac{1}{10^n}$ with the required accuracy.

13.5.2. The test of Dirichlet-Abel. To establish one more refined test for convergence of series, let us derive an interesting identity which is an analogue of the formula of integration by parts. Assume that $u_1, u_2, u_3, \ldots, v_1, v_2, v_3, \ldots$ are arbitrary numbers and $S_n = u_1 + u_2 + \ldots + u_n$, n and p being arbitrary numbers. Then the following identity holds true:

$$\sum_{k=n}^{n+p} u_k v_k = \sum_{k=n}^{n+p-1} S_k (v_k - v_{k+1}) + S_{n+p} v_{n+p} - S_{n-1} v_n. \tag{13.82}$$

Identity (13.82) is usually called *Abel's identity**.

---

* If we rewrite equality (13.82) in the form

$$\sum_{k=n}^{n+p} v_k (S_k - S_{k-1}) = S_{n+p} v_{n+p} - S_{n-1} v_n - \sum_{k=n}^{n+p-1} S_k (v_{k+1} - v_k),$$

it becomes evident that the Abel transformation is actually the formula of summation by parts which is a difference analogue of the formula of integration by parts.

Derivation of Abel's identity. Let us take into account that $u_k = S_k - S_{k-1}$ and substitute this value of u_k into the left-hand side of (13.82). We shall obtain

$$\sum_{k=n}^{n+p} u_k v_k = \sum_{k=n}^{n+p} S_k v_k - \sum_{k=n}^{n+p} S_{k-1} v_k.$$

Let us then reduce the index of summation k in the last sum by unity. We obtain

$$\sum_{k=n}^{n+p} u_k v_k = \sum_{k=n}^{n+p} S_k v_k - \sum_{k=n-1}^{n+p-1} S_k v_{k+1} =$$

$$= \sum_{k=n}^{n+p-1} S_k v_k + S_{n+p} v_{n+p} - \sum_{k=n}^{n+p-1} S_k v_{k+1} - S_{n-1} v_n =$$

$$= \sum_{k=n}^{n+p-1} S_k (v_k - v_{k+1}) + S_{n+p} v_{n+p} - S_{n-1} v_n.$$

We have obtained an expression coinciding with the right-hand side of (13.82) and thus proved Abel's identity.

Theorem 13.15 (Dirichlet-Abel's test). *Suppose we are given a series*

$$\sum_{k=1}^{\infty} u_k v_k. \tag{13.83}$$

The series converges if the following two conditions are complied with:

(1) *the sequence* $\{v_k\}$ *is nonincreasing and infinitesimal;*

(2) *the series* $\sum_{k=1}^{\infty} u_k$ *possesses a bounded sequence of partial sums.*

Proof. Let us denote by S_n the nth partial sum of the series $\sum_{k=1}^{\infty} u_k$. By the hypothesis, there exists a number $M > 0$ such that $|S_n| \leqslant M$ for all the numbers n. By virtue of Cauchy's condition, it is sufficient to prove that for any number $\varepsilon > 0$ there is a number N such that for $n \geqslant N$ and for any natural p

$$\left|\sum_{k=n}^{n+p} u_k v_k\right| < \varepsilon. \tag{13.84}$$

Suppose we are given any $\varepsilon > 0$. Since the sequence $\{v_k\}$ is infinitesimal and nonincreasing, there is a number N, for the positive number $\varepsilon/(2M)$, such that

$$0 \leqslant v_n < \frac{\varepsilon}{2M} \quad (\text{for } n \geqslant N). \tag{13.85}$$

Let us now apply Abel's identity (13.82) to evaluate the quantity on the left-hand side of (13.84). Taking into account that the abso-

lute value of the sum of several quantities does not exceed the sum of their absolute values, that the absolute value of the product is equal to the product of the absolute values and that $v_k \geqslant v_{k+1}$, we obtain

$$\left|\sum_{k=n}^{n+p} u_k v_k\right| \leqslant \sum_{k=n}^{n+p-1} |S_k|(v_k - v_{k+1}) + |S_{n+p}| v_{n+p} + |S_{n-1}| v_n. \tag{13.86}$$

On the right-hand side of (13.86) we make use of the inequality $|S_n| \leqslant M$, valid *for all* the numbers n. We obtain

$$\left|\sum_{k=n}^{n+p} u_k v_k\right| \leqslant M\left\{\sum_{k=n}^{n+p-1}(v_k - v_{k+1}) + v_{n+p}\right\} + Mv_n. \tag{13.87}$$

Next we make note of the fact that the sum in braces is strictly equal to v_n. In that case inequality (13.87) assumes the form

$$\left|\sum_{k=n}^{n+p} u_k v_k\right| \leqslant 2Mv_n. \tag{13.88}$$

If we now apply inequality (13.85) to the right-hand side of (13.88), we shall find that inequality (13.84) is valid for $n \geqslant N$ and for any natural p. We have proved the theorem.

Remark. Theorem 13.14 (Leibniz's test) is a particular case of Theorem 13.15 for* $u_k = (-1)^{k-1}$.

Examples. 1. Find whether the following series is convergent:

$$1 + \frac{1}{2} - \frac{2}{3} + \frac{1}{4} + \frac{1}{5} - \frac{2}{6} + \ldots + \frac{1}{3n-2} + \frac{1}{3n-1} - \frac{2}{3n} + \ldots$$

This series can be regarded as a series of the form of (13.83) for

$$v_k = \frac{1}{k},\ u_1 = 1,\ u_2 = 1,\ u_3 = -2,\ u_4 = 1,\ u_5 = 1,\ u_6 = -2,\ \ldots$$

It is evident that (1) the series $\sum_{k=1}^{\infty} u_k$ possesses a *bounded* sequence of partial sums: $S_1 = 1$, $S_2 = 2$, $S_3 = 0$, $S_4 = 1$, $S_5 = 2$, $S_6 = 0$, . . .; (2) the sequence $\{v_k\}$ is nonincreasing infinitesimal. In accordance with Theorem 13.15 the series is convergent.

2. Elucidate the question concerning the convergence of the series $\sum_{k=1}^{\infty} \frac{\cos kx}{k}$, where x is some fixed real number. Making

* It is evident that the series $\sum_{k=1}^{\infty} u_k = \sum_{k=1}^{\infty} (-1)^{k-1} = 1 - 1 + 1 - 1 + \ldots$ possesses a bounded sequence of partial sums

use of the notation of Theorem 13.15, put $u_k = \cos kx$, $v_k = \frac{1}{k}$ and then evaluate the sequence of the partial sums S_n of the series $\sum_{k=1}^{\infty} u_k$. Since for any number k

$$\sin\left(k+\frac{1}{2}\right)x - \sin\left(k-\frac{1}{2}\right)x = 2\sin\frac{x}{2}\cos kx,$$

we can add up this relation with respect to k from 1 to n and obtain

$$\sin\left(n+\frac{1}{2}\right)x - \sin\frac{1}{2}x = 2\sin\frac{x}{2}\sum_{k=1}^{n}\cos kx = 2S_n\sin\frac{x}{2},$$

whence it follows that

$$S_n = \frac{\sin\left(n+\frac{1}{2}\right)x - \sin\frac{x}{2}}{2\sin\frac{x}{2}}.$$

Thus, for any x which is *not a multiple* of 2π, the sequence of the partial sums S_n is bounded:

$$|S_n| \leqslant \frac{1}{\left|\sin\frac{x}{2}\right|}.$$

In accordance with Theorem 13.15, the series being considered *converges for any value of x which is not a multiple of* 2π. If, now, x *is a multiple of* 2π, then the series under consideration turns into a *harmonic* series and *diverges* as was proved above.

13.6. INFINITE PRODUCTS

13.6.1. Basic notions. The notion of a number series is close in its meaning to the notion of a *continued (infinite) number product.* Suppose we are given an infinite number sequence $v_1, v_2, \ldots, v_k, \ldots$. The formalized expression of the kind of

$$v_1 v_2 v_3 \ldots v_k \ldots = \prod_{k=1}^{\infty} v_k \tag{13.89}$$

is known as a *continued (finite) product.* Separate elements v_k are called the factors of the given infinite product. The product of the first n factors of the given infinite product is usually called the nth partial product and designated as P_n:

$$P_n = v_1 v_2 \ldots v_n = \prod_{k=1}^{n} v_k.$$

Continued product (13.89) is called *convergent* if the sequence of the partial products P_n has a finite limit P *different* from zero*. In the case of convergence of continued product (13.89), the indicated limit P is called the *value of that continued product*, that is, it is written as

$$P = \prod_{k=1}^{\infty} v_k. \tag{13.90}$$

It should be emphasized that equality (13.90) is meaningful only for a convergent continued product. It is clear that the discussion of continued product is in essense a new form of investigation of number sequences, for every given continued product is in a one-to-one correspondence with the sequence of its partial products and every number sequence $\{P_k\}$, whose all terms are different from zero, is in a one-to-one correspondence with a continued product for which this sequence is the sequence of the partial products (it is sufficient to put the factors of this continued product equal to $v_k = P_k/P_{k-1}$ for $k > 1$ and $v_1 = P_1$).

Theorem 13.16. *A necessary condition of convergence of continued product* (13.89) *is the tending of its kth factor to unity as* $k \to \infty$.

Proof. Assume that continued product (13.89) converges and has the value P different from zero. Then $\lim\limits_{k\to\infty} P_{k-1} = \lim\limits_{k\to\infty} P_k = P \neq 0$. Since $v_k = P_k/P_{k-1}$, it follows that $\lim\limits_{k\to\infty} v_k$ exists and is equal to unity.

Note that the convergence of a continued product *is not affected* by the deletion of *any finite number* of the factors of that product (if, of course, there are no factors equal to zero). Since according to the definition accepted above the continued product in which even *one factor* is equal to zero is considered to be *divergent*, we shall *exclude from the consideration, that will follow, continued products in which at least one factor is equal to zero.*

Examples of continued products.

$$1.\ \prod_{k=1}^{\infty} \cos\frac{x}{2^k} = \cos\frac{x}{2}\cos\frac{x}{4}\ \ldots\ \cos\frac{x}{2^k}\ \ldots \tag{13.91}$$

(x is an arbitrary fixed number).

Let us prove that continued product (13.91) converges and has the value $\frac{\sin x}{x}$. We shall compute the nth partial product

$$P_n = \cos\frac{x}{2}\cos\frac{x}{2^2}\ \ldots\ \cos\frac{x}{2^n}. \tag{13.92}$$

* The fact that for $P = 0$ a continued product is usually assumed to be *divergent* is of a conditional character, but all the same, as it will be shown later, this makes it possible to draw an analogy between the convergence of series and that of continued products.

Multiplying both sides of (13.92) by $\sin(x/2^n)$ and applying successively the formula for the sine of the double angle $\sin 2y = 2 \sin y \cos y$, we obtain

$$P_n \sin\frac{x}{2^n} = \frac{1}{2^n} \sin x.$$

The last formula yields*

$$P_n = \frac{\sin x}{x} \left\{ \frac{\left(\frac{x}{2^n}\right)}{\sin\left(\frac{x}{2^n}\right)} \right\}.$$

Since the expression in braces tends to unity as $n \to \infty$ (by virtue of the first remarkable limit), it follows that $\lim\limits_{n\to\infty} P_n$ exists and is equal to $\frac{\sin x}{x}$. Thus we have proved that continued product (13.91) converges and has the value $\frac{\sin x}{x}$.

$$2. \prod_{k=2}^{\infty} \left[1 - \frac{2}{k(k+1)}\right] = \prod_{k=2}^{\infty} \frac{(k-1)(k+2)}{k(k+1)} =$$

$$= \frac{1}{2} \cdot \frac{4}{3} \cdot \frac{2}{3} \cdot \frac{5}{4} \cdots \frac{(k-1)}{k} \cdot \frac{(k+2)}{(k+1)} \cdots \qquad (13.93)$$

Let us prove that continued product (13.93) converges and has the value 1/3. We shall calculate the partial product P_n:

$$P_n = \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{4} \cdots \frac{n-1}{n} \cdot \frac{4}{3} \cdot \frac{5}{4} \cdot \frac{6}{5} \cdots \frac{n+2}{n+1} = \frac{1}{n} \cdot \frac{v+2}{3}.$$

Now it is evident that $\lim\limits_{n\to\infty} P_n = \lim\limits_{n\to\infty} \frac{n+2}{3n}$ exists and is equal to $\frac{1}{3}$.

13.6.2. Connection between the convergence of continued products and series. If continued product (13.89) is convergent, then, by virtue of Theorem 13.16, all its terms v_k, beginning with some number k, are positive**. Inasmuch as the finite number of the first terms does not influence the convergence of a continued product at all, it follows that when studying convergence of continued products, we can, without loss of generality, consider only those continued products whose *all terms are positive.*

---

* We assume that $x \neq 0$. If $x = 0$, then all the terms of (13.91) and its value are equal to unity.

** Since $\lim\limits_{k\to\infty} v_k = 1$.

Theorem 13.17. *For the positive continued product* (13.89) *to converge, it is necessary and sufficient for the following series to converge:*

$$\sum_{k=1}^{\infty} \ln v_k. \tag{13.94}$$

In case it is convergent, the sum S of series (13.94) *and the value P of product* (13.89) *are connected by the formula*

$$P = e^S. \tag{13.95}$$

Proof. Having denoted by P_n the n-th partial product of continued product (13.89) and by S_n, the n-th partial sum of series (13.94), we can wri e

$$S_n = \ln P_n, \quad P_n = e^{S_n}.$$

By virtue of continuity of an exponential function for all the values of the argument and by virtue of continuity of a logarithmic function for all the positive values of the argument, the sequence P_n converges if and only if the n-th partial sum S_n converges, and if $\lim\limits_{n\to\infty} S_n = S$, then $\lim\limits_{n\to\infty} P_n = e^S$. We have proved the theorem.

When we investigate a continued product as to its convergence, it turns out to be most convenient to represent it in the form

$$\prod_{k=1}^{\infty} (1+u_k) = (1+u_1)(1+u_2) \ldots (1+u_k) \ldots \tag{13.96}$$

In this case, in accordance with the assumption accepted above, we consider, of course, all $u_k > -1$.

Theorem 13.17 asserts that the question concerning the convergence of product (13.96) is equivalent to that concerning the convergence of the series

$$\sum_{k=1}^{\infty} \ln (1+u_k). \tag{13.97}$$

We can prove now one more assertion.

Theorem 13.18. *If all u_k (at least beginning with some number k) retain sign, then for continued product* (13.96) *to be convergent, it is necessary and sufficient for the following series to be convergent:*

$$\sum_{k=1}^{\infty} u_k. \tag{13.98}$$

Proof. Since the condition $\lim\limits_{k\to\infty} u_k = 0$ is necessary both for the convergence of series (13.98) and for the convergence of product (13.96), we can assume the condition to be fulfilled both in proving the ne-

cessity and in proving the sufficiency. But it follows from the indicated condition and from the asymptotic formula*

$$\ln(1+y)=y+o(y)$$

that

$$\lim_{k\to\infty}\frac{\ln(1+u_k)}{u_k}=1 \tag{13.99}$$

and

$$\lim_{k\to\infty}\frac{u_k}{\ln(1+u_k)}=1. \tag{13.100}$$

Since by the hypothesis of the theorem all the terms of series (13.97) and (13.98) retain sign beginning with some number k, conditions (13.99) and (13.100) make it possible to assert, by virtue of the corollary of Comparison Theorem 13.3, that series (13.98) converges if and only if series (13.97) is convergent. We have proved the theorem.

Examples. (1) The divergence of a harmonic series and Theorem 13.18 yield the divergence of the following continued products:

$$\prod_{k=1}^{\infty}\left(1+\frac{1}{k}\right)=(1+1)\left(1+\frac{1}{2}\right)\left(1+\frac{1}{3}\right)\cdots\left(1+\frac{1}{k}\right)\cdots,$$

$$\prod_{k=1}^{\infty}\left(1-\frac{1}{k+1}\right)=\left(1-\frac{1}{2}\right)\left(1-\frac{1}{3}\right)\cdots\left(1-\frac{1}{k+1}\right)\cdots$$

It is easy to realize that the first of these products diverges to $+\infty$ and the second to zero.

(2) The same Theorem 13.18 and the convergence of series (13.33) at $\alpha>1$ yield the convergence, at $\alpha>1$, of the following continued products:

$$\prod_{k=1}^{\infty}\left(1+\frac{1}{k^{\alpha}}\right)=(1+1)\left(1+\frac{1}{2^{\alpha}}\right)\left(1+\frac{1}{3^{\alpha}}\right)\cdots\left(\frac{1}{k^{\alpha}}\right)\cdots,$$

$$\prod_{k=1}^{\infty}\left[1-\frac{1}{(k+1)^{\alpha}}\right]=\left(1-\frac{1}{2^{\alpha}}\right)\left(1-\frac{1}{3^{\alpha}}\right)\cdots\left(1-\frac{1}{(k+1)^{\alpha}}\right)\cdots$$

The notion of *absolute* and *conditional* convergence is introduced for continued products by analogy with series. Continued product (13.96) is said to be *absolutely convergent* if and only if series (13.97) converges absolutely. Theorems of Cauchy 13.11 and Riemann 13.10 make it possible to infer that an absolutely convergent product possesses

* See 4.7.

commutativity whereas a conditionally convergent product obviously does not possess it.

The following assertion is valid.

Theorem 13.19. *Continued product* (13.96) *is absolutely convergent if and only if series* (13.98) *is absolutely convergent.* To prove this theorem, it is sufficient to show that the series $\sum_{k=1}^{\infty} |u_k|$ converges if and only if the series $\sum_{k=1}^{\infty} |\ln(1+u_k)|$ converges. This last convergence easily follows from the existence of limits (13.99) and (13.100). The details of the argument are left to the reader.

In conclusion we shall consider some more examples.

1°. Let us consider the continued product

$$x\prod_{k=1}^{\infty}\left(1-\frac{x^2}{k^2\pi^2}\right)=x\left(1-\frac{x^2}{\pi^2}\right)\left(1-\frac{x^2}{2^2\pi^2}\right)\left(1-\frac{x^2}{3^2\pi^2}\right)\ldots \quad (13.101)$$

The series $\sum_{k=1}^{\infty}\frac{1}{k^2}$ being convergent, it follows, in accordance with Theorems 13.18 and 13.19, that continued product (13.101) is absolutely convergent for any given value of x different from $l\pi$ (where $l=0, \pm 1, \ldots$). In Supplement 2 to this chapter we shall prove that this product converges to the value $\sin x$. That will be the substantiation of the expansion of the function $\sin x$ into the continued product

$$\sin x=x\prod_{k=1}^{\infty}\left(1-\frac{x^2}{k^2\pi^2}\right). \quad (13.102)$$

2°. The following expansion can be easily obtained from expansion (13.102) by the use of the relation $\cos x=\frac{\sin 2x}{2\sin x}$:

$$\cos x=\prod_{k=1}^{\infty}\left[1-\frac{4x^2}{(2k-1)^2\pi^2}\right]. \quad (13.103)$$

The absolute convergence of the product appearing on the right-hand side of (13.103) for any x different from $\frac{\pi}{2}(2l-1)$ $(l=0, \pm 1, \ldots)$ follows from Theorems 13.18 and 13.19 and from the convergence of the series $\sum_{k=1}^{\infty}\frac{1}{(2k-1)^2}$.

3°. Putting $x=\frac{\pi}{2}$ in expansion (13.102), we obtain

$$\frac{2}{\pi}=\prod_{k=1}^{\infty}\left(1-\frac{1}{4k^2}\right)=\prod_{k-1}^{\infty}\frac{4k2-1}{4k_2}=\prod_{k=1}^{\infty}\frac{(2k-1)(2k+1)}{(2k)^2}$$

whence follows the so-called *Wallis's** formula*

$$\frac{\pi}{2}=\prod_{k=1}^{\infty}\frac{(2k)^2}{(2k-1)(2k+1)}=\frac{2}{1}\cdot\frac{2}{3}\cdot\frac{4}{3}\cdot\frac{4}{5}\cdot\ldots\cdot\frac{2k}{2k-1}\cdot\frac{2k}{2k+1}\cdot\ldots \tag{13.104}$$

By means of easy transformations Wallis's formula can be reduced to the form

$$\frac{\pi}{2}=\lim_{k\to\infty}\frac{1}{2k+1}\left[\frac{2^{2k}(k!)^2}{(2k)!}\right]^2. \tag{13.104*}$$

Originally Wallis's formula was used for approximate calculations of the number π. At present there are more efficient methods for calculating the number π. Wallis's formula (13.104) may be of interest for certain theoretical investigations**.

SUPPLEMENT 1 TO CHAPTER 13

AUXILIARY THEOREM TO 13.2.3

Theorem 13.20. *Suppose p_k are arbitrary positive numbers. Then, if there exists a limit*

$$\lim_{k\to\infty}\frac{p_{k+1}}{p_k}=L, \tag{13.105}$$

then there also exists a limit $\lim\limits_{k\to\infty}\sqrt[k]{p_k}$, with the formula

$$\lim_{k\to\infty}\sqrt[k]{p_k}=\lim_{k\to\infty}\frac{p_{k+1}}{p_k}=L \tag{13.106}$$

being valid.

Proof. Let us first prove the following auxiliary assertion***: *if the sequence of the positive numbers $a_1, a_2, \ldots, a_k, \ldots$ converges to some number L, then the sequence of the geometric mean of these numbers $b_k=\sqrt[k]{a_1a_2\ldots a_k}$ converges to the same number L.* To prove the auxiliary assertion, let us make a note of the fact that by virtue of the continuity of a logarithmic function, $\lim\limits_{k\to\infty}\ln a_k=\ln L$ for $L>0$. (The last equality is formally valid for $L=0$ as well when $\ln L=-\infty$.) But then, according to the theorem on the limit of the arithmetic mean (see the supplement to Chap. 3, Example 1) there is a limit

$$\lim_{k\to\infty}\ln\sqrt[k]{a_1a_2\ldots a_k}=\lim_{k\to\infty}\frac{\ln a_1+\ln a_2+\ldots+\ln a_k}{k}=\ln L.$$

---

* John Wallis (1616-1703), a British mathematician.

** In particular, it can be employed to establish the so-called Stirling's formula (see Part 2 of the present course). James Stirling (1692-1770), a British mathematician.

*** It should be pointed out that the assertion itself may also be of interest.

(The last equality is valid for $L = 0$ as well if $\ln L = -\infty$.) By virtue of the continuity of an exponential function, we shall obtain from the last equality

$$\lim_{k\to\infty} \sqrt[k]{a_1 a_2 \dots a_k} = \lim_{k\to\infty} e^{\ln \sqrt[k]{a_1 a_2 \dots a_k}} = e^{\ln L} = L.$$

(The argument is valid for $L = 0$ as well.)

We have proved the auxiliary assertion. Applying this assertion to the numbers $a_1 = p_1$, $a_2 = \frac{p_2}{p_1}$, $a_3 = \frac{p_3}{p_2}$, ..., $a_k = \frac{p_k}{p_{k-1}}$, ..., we shall establish the existence of the limit $\lim_{k\to\infty} \sqrt[k]{p_k}$ and equality (13.106). The proof of Theorem 13.20 is complete.

SUPPLEMENT 2 TO CHAPTER 13

EXPANSION OF THE FUNCTION sin x INTO A CONTINUED PRODUCT

For the sake of convenience we shall divide the derivation of formula (13.102) into stages.

1°. Assume m to be *any positive odd* number: $m = 2n + 1$. First of all we shall prove that for any value θ* different from $k\pi$ ($k = 0, \pm 1, \dots$), the following formula holds true:

$$\frac{\sin m\theta}{m \sin \theta} = \left(1 - \frac{\sin^2 \theta}{\sin^2 \frac{\pi}{m}}\right)\left(1 - \frac{\sin^2 \theta}{\sin^2 \frac{2\pi}{m}}\right) \dots$$

$$\dots \left(1 - \frac{\sin^2 \theta}{\sin^2 \frac{n\pi}{m}}\right), \quad n = \frac{m-1}{2}. \tag{13.107}$$

To establish formula (13.107), we shall proceed from de Moivre's formula (see 7.1)

$$\cos m\theta + i \sin m\theta = (\cos \theta + i \sin \theta)^m$$

Writing out the right-hand side of this formula with the aid of Newton's binomial and comparing the imaginary parts, we obtain

$$\sin m\theta = m \cos^{m-1} \theta \sin \theta - \frac{m(m-1)(m-2)}{1 \cdot 2 \cdot 3} \cos^{m-3} \theta \sin^3 \theta + \dots$$

Taking into account that $m = 2n + 1$, we shall have

$$\frac{\sin m\theta}{m \sin \theta} = \cos^{2n} \theta - \frac{(m-1)(m-2)}{1 \cdot 2 \cdot 3} \cos^{2n-2} \theta \sin^2 \theta + \dots \tag{13.108}$$

All the exponents in cosines and sines on the right-hand side of (13.108) are *even* so that if we replace $\cos^2 \theta$ by $1 - \sin^2 \theta$, we obtain, *on the right-hand side of* (13.108), *a polynomial of degree n with respect to* $\sin^2 \theta$. Having put $z = \sin^2 \theta$,

---

* In what follows we shall be interested only in the values of θ in the interval $0 < |\theta| < \pi$.

we shall designate this polynomial as $F(z)$ and its roots as $\alpha_1, \alpha_2, \ldots, \alpha_n$. Since $z = \sin^2\theta \to 0$ as $\theta \to 0$ and inasmuch as the left-hand side of (13.108) tends to unity as $\theta \to 0$, the polynomial $F(z)$ can be represented as

$$\frac{\sin m\theta}{m \sin\theta} = E(z) = \left(1 - \frac{z}{\alpha_1}\right)\left(1 - \frac{z}{\alpha_2}\right) \ldots \left(1 - \frac{z}{\alpha_n}\right).$$

It remains to determine the roots $\alpha_1, \alpha_2, \ldots, \alpha_n$. Noting that these roots correspond to the zeros of the function $\sin m\theta$, we obtain

$$\alpha_1 = \sin^2\frac{\pi}{m}, \quad \alpha_2 = \sin^2\frac{2\pi}{m}, \quad \ldots, \quad \alpha_n = \sin^2\frac{n\pi}{m}.$$

Thus we have established formula (13.107).

2°. Having put $\theta = \frac{x}{m}$ in formula (13.107) and having assumed that $0 < |x| < \pi m$, we shall represent this formula in the form

$$\frac{\sin x}{m \sin\frac{x}{m}} = \prod_{k=1}^{n}\left(1 - \frac{\sin^2\frac{x}{m}}{\sin^2\frac{k\pi}{m}}\right). \tag{13.109}$$

Let us fix *any* (nonzero) value of x and take two arbitrary natural numbers p and n complying with the inequalities $2\frac{|x|}{\pi} < p < n = \frac{m-1}{2}$. Then formula (13.109) can be written in the form

$$\frac{\sin x}{m \sin\frac{x}{m}} = \prod_{k=1}^{p}\left(1 - \frac{\sin^2\frac{x}{m}}{\sin^2\frac{k\pi}{m}}\right) R_p(x), \tag{13.110}$$

where

$$R_p(x) = \prod_{k=p+1}^{n}\left(1 - \frac{\sin^2\frac{x}{m}}{\sin^2\frac{k\pi}{m}}\right). \tag{13.111}$$

Let us first of all estimate $R_p(x)$. Since $2\frac{|x|}{\pi} < p < n = \frac{m-1}{2}$, the arguments of all the sines appearing in (13.111) belong to the interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. It is clear, moreover, that for all k appearing in this formula $|x| < \frac{k\pi}{2}$ and, consequently,

$$0 < \frac{\sin^2\frac{x}{m}}{\sin^2\frac{k\pi}{m}} < \frac{\sin^2\frac{k\pi}{2m}}{\sin^2\frac{k\pi}{m}} = \frac{1}{4\cos^2\frac{k\pi}{2m}} < \frac{1}{2}$$

$\left(\text{since } \frac{k\pi}{m} < \frac{\pi}{2}, \text{ i.e. } \frac{k\pi}{2m} < \frac{\pi}{2} \text{ and, therefore, } \cos^2\frac{k\pi}{2m} > \frac{1}{2}\right)$.

Since the inequalities $1 > 1 - \beta > e^{-2\beta}$* hold true for any β from the interval $0 < \beta < \frac{1}{2}$, the following formula is valid for all numbers k exceeding p:

$$1 > 1 - \frac{\sin^2 \frac{x}{m}}{\sin^2 \frac{k\pi}{m}} > e^{-2 \cdot \frac{\sin^2 \frac{x}{m}}{\sin^2 \frac{k\pi}{m}}}. \tag{13.112}$$

Multiplying term-by-term inequalities (13.112) written for the values $k = p + 1, p + 2, \ldots, n$, we shall obtain the following estimation for $R_p(x)$:

$$1 > R_p(x) > e^{-2 \sin^2 \frac{x}{m} \sum_{k=p+1}^{n} \frac{1}{\sin^2 \frac{k\pi}{m}}}. \tag{13.113}$$

Bearing in mind that the argument $\frac{k\pi}{m}$ is in the first quarter and that $1 \geqslant \frac{\sin \beta}{\beta} \geqslant \frac{2}{\pi}$** for any β from the first quarter, we obtain

$$\frac{1}{\sin^2 \frac{k\pi}{m}} < \frac{1}{\left(\frac{2}{\pi}\right)^2 \left(\frac{k\pi}{m}\right)^2} = \frac{m^2}{4k^2} < \frac{m^2}{4}\left[\frac{1}{k-1} - \frac{1}{k}\right].$$

Thus we have

$$e^{-2 \sin^2 \frac{x}{m} \sum_{k=p+1}^{n} \frac{1}{\sin^2 \frac{k\pi}{m}}} > e^{-\frac{m^2}{2} \sin^2 \frac{x}{m} \sum_{k=p+1}^{\infty} \left[\frac{1}{k-1} - \frac{1}{k}\right]} =$$

$$= e^{-\frac{m^2}{2p} \sin^2 \frac{x}{m}}.$$

The last inequality makes it possible to strengthen evaluation (13.113) as follows:

$$1 > R_p(x) > e^{-\frac{m^2}{2p} \sin^2 \frac{x}{m}}. \tag{13.114}$$

---

* The right-hand inequality can be easily obtained from Maclaurin's formula $e^{-2\beta} = 1 - 2\beta + \frac{(2\beta)^2}{2} - \ldots < 1 - 2\beta + 2\beta^2 < 1 - \beta$, since $2\beta^2 < \beta$.

** These inequalities follow from the fact that the relation $\frac{\sin \beta}{\beta}$ decreases from 1 to $\frac{2}{\pi}$ upon the change of β from 0 to $\frac{\pi}{2}$. In its turn, the decrease of the function $\frac{\sin \beta}{\beta}$ follows from the fact that $\left(\frac{\sin \beta}{\beta}\right)' = \frac{\cos \beta}{\beta^2}(\beta - \tan \beta) < 0$ everywhere on the interval $0 < \beta < \frac{\pi}{2}$.

3°. Let us make the number m in formula (13.110) tend to infinity retaining fixed the value x and the number p. Since $\lim\limits_{m\to\infty} m\sin\frac{x}{m}=x$, $\lim\limits_{m\to\infty} m^2\sin^2\frac{k\pi}{m}=k^2\pi^2$, there is a limit of the left-hand side of (13.110) equal to $\frac{\sin x}{x}$ and a limit of a finite product $\prod\limits_{k=1}^{p}\left(1-\frac{\sin^2\frac{x}{m}}{\sin^2\frac{k\pi}{m}}\right)$ equal to $\prod\limits_{k=1}^{p}\left(1-\frac{x^2}{k^2\pi^2}\right)$. We shall also assume that the last limit is different from zero, since when it is equal to zero, $\sin x=0$ and expansion (13.102) is established. But then there also exists a limit $\lim\limits_{m\to\infty} R_p(x)$. Let us denote this limit by $\hat{R}_p(x)$. Inequalities (13.114), valid for any number m, and Theorem 3.13 yield

$$1\geqslant\hat{R}_p(x)\geqslant e^{-\frac{x^2}{2p}}. \tag{13.115}$$

In the limit, as $m\to\infty$, formula (13.110) yields

$$\frac{\sin x}{x}=\prod_{k=1}^{p}\left(1-\frac{x^2}{k^2\pi^2}\right)\hat{R}_p(x). \tag{13.116}$$

4°. It remains, retaining fixed x, to make the number p tend to infinity in formula (13.116). Since the left-hand side of (13.116) does not depend on p and the limit $\lim\limits_{p\to\infty}\hat{R}_p(x)$, by virtue of inequalities (13.115) and Theorem 3.14, exists and is equal to unity, there exists a limit

$$\lim_{p\to\infty}\prod_{k=1}^{p}\left(1-\frac{x^2}{k^2\pi^2}\right)=\frac{\sin x}{x}.$$

Thus we have determined the expansion for $\sin x$ (13.102).

Remark. By a complete analogy with expansions (13.102) for $\sin x$ and (13.103) for $\cos x$, it is possible to obtain the *expansion into continued products of the hyperbolic functions*

$$\sinh x=x\prod_{k=1}^{\infty}\left(1+\frac{x^2}{k^2\pi^2}\right) \text{ and } \cosh x=\prod_{k=1}^{\infty}\left[1+\frac{4x^2}{(2k-1)^2\pi^2}\right].$$

Note that the expansions for $\sin x$, $\cos x$, $\sinh x$, $\cosh x$ immediately yield the expansions into continued products of the functions $\tan x$, $\cot x$, $\tanh x$ and $\coth x$.

SUPPLEMENT 3 TO CHAPTER 13

GENERALIZED METHODS OF SUMMATION OF DIVERGENT SERIES

Throughout Chapter 13 we used the term the sum of the series

$$\sum_{k=1}^{\infty} u_k = u_1 + u_2 + u_3 + \ldots + u_k + \ldots \tag{13.117}$$

for the limit S of the sequence $\{S_n\}$ of the partial sums of that series (under the condition that the limit does exist).

In a number of problems of mathematical analysis, both of theoretical and practical interest, we have to operate with series whose sequence of partial sums does not converge and *there is no sum in the ordinary meaning indicated in Chapter 13*. The problem naturally arises *of generalizing the concept of the sum of a series and of summation* of series (13.117), divergent in the ordinary meaning, *with the aid of some generalized methods*. In the present supplement we shall discuss some generalized methods of summation of divergent series.

First of all we shall give a general characterization of the methods of summation we shall deal with. It is reasonable to require that the generalized concept of the sum *should include* the ordinary meaning of the sum. To be more exact, *a series converging in the ordinary sense and possessing an ordinary sum S must possess a generalized sum, this sum being also equal to S*. The method of summation possessing the indicated property is called *regular* or *consistent*.

Next it is natural to make the generalized sum comply with the following condition: *if the series* $\sum_{k=1}^{\infty} u_k$ *possesses a generalized sum U and the series* $\sum_{k=1}^{\infty} v_k$ *possesses a generalized sum V, then the series* $\sum_{k=1}^{\infty} (Au_k + Bv_k)$, *where A and B are arbitrary constants, possesses a generalized sum* $(AU + BV)$. The method of summation complying with the indicated condition is called *linear*. In the analysis and its applications use is usually made *only of regular linear summation methods*. We shall discuss here two methods of generalized summation which are of especial interest for applications.

3.1. Cesàro's* method (or method of arithmetic means). *We say that we sum up series* (13.117) *by the Cesàro method if there exists a limit of the arithmetic mean of the partial sums of that series*

$$\lim_{n\to\infty} \frac{S_1 + S_2 + \ldots + S_n}{n}. \tag{13.118}$$

In this case the limit (13.118) *is called the generalized, in the sense of Cesàro, sum of series* (13.117).

The *linearity* of Cesàro's summation method is obvious. Its *regularity* follows from Example 1 considered in Supplement 1 to Chap. 3. Indeed, it follows from the indicated example that if the sequence $\{S_n\}$ of the partial sums of series (13.117) converges to the number S, then the limit (13.118) exists and is also equal to S.

We shall now give some examples of the series which cannot converge in the ordinary sense but can be summed up by the Cesàro method.

---

* Ernesto Cesàro (1859-1906), an Italian mathematician.

(1) Let us consider the obviously divergent series

$$\sum_{k=1}^{\infty} (-1)^{k-1} = 1 - 1 + 1 - 1 + \dots$$

Since all the *even* partial sums S_{2n} of this series are equal to zero, and all the *odd* partial sums S_{2n-1} are equal to unity, the limit (13.118) exists and is equal to $\frac{1}{2}$. Thus the series being considered can be summed up by the use of Cesàro's method and its sum in the sense of Cesàro is equal to $\frac{1}{2}$.

(2) Assuming that x is any fixed real number belonging to the interval $0 < x < 2\pi$, we shall consider the obviously divergent* series

$$\sum_{k=1}^{\infty} \cos kx = \cos x + \cos 2x + \cos 3x + \dots \qquad (13.119)$$

The partial sum S_n of this series has been estimated in Example 2 at the end of 13.5:

$$S_n = \frac{\sin\left(n + \frac{1}{2}\right)x - \sin\frac{x}{2}}{2\sin\frac{x}{2}}.$$

Let us calculate the arithmetic mean of the partial sums:

$$\frac{S_1 + S_2 + \dots + S_n}{n} = \frac{1}{2n\sin\frac{x}{2}}\left[\sum_{m=1}^{n} \sin\left(m + \frac{1}{2}\right)x\right] - \frac{1}{2} =$$

$$= \frac{1}{4n\sin^2\frac{x}{2}}\left[\sum_{m=1}^{n} (\cos mx - \cos(m+1)x)\right] - \frac{1}{2} =$$

$$= \frac{\cos x - \cos(n+1)x}{4n\sin^2\frac{x}{2}} - \frac{1}{2}.$$

From this it is evident that

$$\lim_{n\to\infty} \frac{S_1 + S_2 + \dots + S_n}{n} = -\frac{1}{2}.$$

Thus series (13.119) is summable by Cesàro's method, the sum being equal to $\left(-\frac{1}{2}\right)$ in the sense of Cesàro.

3.2. Poisson-Abel's summation method.** This method consists in the following. *A power series*

$$\sum_{k=1}^{\infty} u_k x^{k-1} = u_1 + u_2 x + u_3 x^2 + \dots + u_k x^{k-1} + \dots \qquad (13.120)$$

* The divergence of series (13.119) can easily be seen from the expression for its partial sum illustrated below.

** Siméon Denis Poisson (1781-1840), a French mathematician.

is constructed from the given series (13.117). *If the indicated power series converges for all* x *from the interval* $0 < x < 1$ *and the sum* $S(x)$ *of this series possesses a left-hand limiting value* $\lim_{x \to 1-0} S(x)$ *at point* $x = 1$, *then series* (13.117) *is said to be summable by Poisson-Abel's method. In this case the indicated limiting value is called the sum of series* (13.117) *in the sense of Poisson-Abel.*

The *linearity* of Poisson-Abel's summation method is undoubtful, the *regularity* of the method, however, should be proved. Assume that series (13.117) converges in the ordinary sense and has a sum equal to S. It is required to prove (1) that series (13.120) converges for any x belonging to the intervall $0 < x < 1$, (2) that the sum $S(x)$ of series (13.120) possesses at point $x = 1$ a left-hand limiting value equal to S.

Let us first prove assertion (1). Since series (13.117) is convergent, the sequence of its terms is *infinitesimal* and, consequently, *bounded*, that is, there is a number M such that for all the numbers k

$$|u_k| \leqslant M. \tag{13.121}$$

Proceeding from inequality (13.121), let us evaluate the absolute value of the k-th term of series (13.120) assuming that x is any number from the interval $0 < x < 1$. We obtain

$$|u_k x^{k-1}| \leqslant M |x|^{k-1}.$$

Since $|x| < 1$, the series $\sum_{k=1}^{\infty} |x|^{k-1}$ converges. Consequently, by virtue of Remark 2 to Comparison Theorem 13.3, series (13.120) converges, too.

Let us now prove assertion (2). Suppose that S_n is the n-th partial sum of series (13.117) and S is its ordinary sum. By means of Abel's* transformation it is easy to verify that for any x from the interval $0 < x < 1$ the following identity holds true:

$$\sum_{k=1}^{\infty} u_k x^{k-1} = (1-x) \sum_{k=1}^{\infty} S_k x^{k-1}. \tag{13.122}$$

The next move is to subtract identity (13.122) from the following obvious identity:

$$S = (1-x) \sum_{k=1}^{\infty} S x^{k-1}.$$

Denoting the k-th remainder of series (13.117) by r_k, we shall have

$$S - \sum_{k=1}^{\infty} u_k x^{k-1} = (1-x) \sum_{k=1}^{\infty} r_k x^{k-1},$$

or

$$S - S(x) = (1-x) \sum_{k=1}^{\infty} r_k x^{k-1}. \tag{13.123}$$

Our aim is to prove that for any $\varepsilon > 0$ there is $\delta > 0$ such that the left-hand side of (13.123) is smaller than ε for all x satisfying the inequalities $1 - \delta < x < 1$. Since the remainder r_k of series (13.117) tends to zero as $k \to \infty$, there is

* Abel's transformation (13.82) was established in Sec. 13.5.2. In the present case it is necessary to put $n = 1$, $S_{n-1} = 0$ in (13.82) and then make p tend to infinity.

a number k_0, for the positive number $\frac{\varepsilon}{2}$, such that $r_k < \frac{\varepsilon}{2}$ for $k \geqslant k_0$. Thus we have

$$\left|(1-x)\sum_{k=k_0}^{\infty} r_k x^{k-1}\right| < \frac{\varepsilon}{2}\left|(1-x)\sum_{k=k_0}^{\infty} x^{k-1}\right| < \frac{\varepsilon}{2}.$$

It remains to prove that for all x sufficiently close to unity

$$\left|(1-x)\sum_{k=1}^{k_0-1} r_k x^{k-1}\right| < \frac{\varepsilon}{2},$$

but this is obvious, for the sum appearing in the last inequality is bounded. We have proved the regularity of Poisson-Abel's method. By way of example we shall again consider a divergent series

$$\sum_{k=1}^{\infty} (-1)^{k-1} = 1-1+1-1+\ldots \tag{13.124}$$

For this series we shall compile a power series of the form of (13.120):

$$\sum_{k=1}^{\infty} (-1)^{k-1} x^{k-1} = 1-x+x^2-x^3+\ldots$$

It is evident that the last series converges for all x belonging to the interval $0 < x < 1$ and has a sum equal to $S(x) = \frac{1}{1+x}$. Since

$$\lim_{x\to 1-0} S(x) = \lim_{x\to 1-0} \frac{1}{1+x} = \frac{1}{2},$$

series (13.124) is summable by Poisson-Abel's method and its sum is equal to $\frac{1}{2}$ in the sense of Poisson-Abel.

Pay attention to the fact that the sum of series (13.124) in Poisson-Abel's sense coincides with its sum in the sense of Cesàro. This is not fortuitous and it can be proved that if a series is summable by Cesàro's method it is also summable by Poisson-Abel's method, the sum of the series in Cesàro's sense coinciding with its sum in the sense of Poisson-Abel. Moreover, there exist series summable by Poisson-Abel's method but not summable by Cesàro's method*. A detailed discussion of various methods of generalized summation of divergent series can be found in *Divergent Series* by G. Hardy, Oxford Press, 1949.

* Thus we can say that Poisson-Abel's method is "stronger" than Cesàro's method of summation.

CHAPTER 14

FUNCTIONS OF SEVERAL VARIABLES

14.1. THE CONCEPT OF A FUNCTION OF SEVERAL VARIABLES

14.1.1. On functional relationships between several variable quantities. When dealing with problems of natural sciences we often come across relationships between several variable quantities such that the values of one of these quantities are completely determined by the values of the other quantities. Thus, when considering physical properties of a body, such as its density ρ or temperature T, for instance, we have to take into account a change in these properties upon transition from one point of the body to another. Since each point of a body is characterized by three Cartesian coordinates x, y and z, the properties in question (density ρ or temperature T) are determined by the values of *three* variables x, y and z.

When we consider physical processes varying in time, the values of the physical properties are determined by the values of *four* variables: three coordinates x, y and z of the point and time t. For instance, when sound vibrations of a gas are studied, its density ρ and pressure p are specified by the values of four variables x, y, z and t. To study relationships of this kind, we introduce in this chapter the concept of a function of several variables and present the means of investigating these functions.

In the theory of functions of several variables it is convenient to use geometric terminology. It is immediately seen that the domain of definition of a function of two (or three) variables is a set of points in a plane (or space). To geometrize our conception of a function of m variables, it is convenient to introduce the notion of an m-dimensional space generalizing the well-known notions of two-dimensional plane and three-dimensional space. We shall begin the exposition with elucidating the necessary geometric concepts.

14.1.2. The concepts of the Euclidean plane and the Euclidean space. The concepts of coordinates of points in a plane and in space and the formula for determining the distance between two points, well-known from analytic geometry, can be used for analytic introduction of the concepts of a plane and space. *A set of various ordered pairs (x, y) of the real numbers x and y is called a* **coordinate plane.**

Every such pair (x, y) will be called a point in that plane and denoted by a single letter M. The numbers x and y are called the coordinates of point M. The notation $M(x, y)$ means that point M has the coordinates x and y.

A coordinate plane is called a **Euclidean plane** *if the distance* $\rho(M', M'')$ *between any two points* $M'(x', y')$ *and* $M''(x'', y'')$ *of the coordinate plane is specified by the formula*

$$\rho(M', M'') = \sqrt{(x''-x')^2+(y''-y')^2}.$$

The concepts of a coordinate space and a Euclidean space can be introduced by a complete analogy. *A set of various ordered triples* (x, y, z) *of the numbers* x, y *and* z *is called a* **coordinate space**. Every triple (x, y, z) will be termed a point of space and denoted by the single letter M. The numbers x, y, z are called the coordinates of point M, and the notation $M(x, y, z)$ means that point M has the coordinates x, y, z.

A coordinate space is termed a **Euclidean space** *if the distance between every two points* $M'(x', y', z')$ *and* $M''(x'', y'', z'')$ *of the coordinate space is specified by the formula*

$$\rho(M', M'') = \sqrt{(x''-x')^2+(y''-y')^2+(z''-z')^2}.$$

The concepts of a coordinate plane and a coordinate space we have introduced are the analogues of a number line, and a Euclidean plane and a Euclidean space are the analogues of a *Euclidean line* which can be defined as a number line the distance $\rho(x', x'')$ between whose any points x' and x'' is determined by the formula $\rho(x', x'') = \sqrt{(x''-x')^2} = |x''-x'|$.

Let us consider certain sets $\{M\}$ of points of a Euclidean plane and Euclidean space.

1°. As is known, the set $\{M\}$ of points of a Euclidean plane whose coordinates x and y satisfy the inequality $(x-a)^2+(y-b)^2 \leqslant R^2$ is called a *circle* of radius R with centre at point $M_0(a, b)$. If the coordinates x and y satisfy the strict inequality $(x-a)^2+(y-b)^2 < R^2$, then the set $\{M\}$ is termed an *open circle*. In a Euclidean space the set $\{M\}$ of points whose coordinates x, y and z comply with the inequality $(x-a)^2+(y-b)^2+(z-c)^2 \leqslant R^2$ is known as a *sphere* of radius R with centre at point $M_0(a, b, c)$. If the coordinates x, y and z satisfy the respective strict inequality, then the set $\{M\}$ is called an *open sphere**.

2°. The set $\{M\}$ of points of a Euclidean plane (Euclidean space) whose coordinates x and y (x, y and z) satisfy the inequalities $|x-a| \leqslant d_1$ and $|y-b| \leqslant d_2$ ($|x-a| \leqslant d_1$, $|y-b| \leqslant d_2$ and $|z-c| \leqslant d_3$), is called a *coordinate rectangle* (*coordinate parallelepiped*) with centre at point $M_0(a, b)$ (at point $M_0(a, b, c)$).

14.1.3. The concept of a function of two or three variables. Making use of geometric terminology, we can reformulate the familiar notion of a function of one variable as follows.

* A circle and a sphere are, evidently, the sets $\{M\}$ of points of a plane and a space for which $\rho(M, M_0) \leqslant R$.

If every point M from a certain set $\{M\}$ of points of a Euclidean line is associated with some number u in accordance with the known law, then a function $u = u(M)$ or $u = f(M)$ is said to be specified on the set $\{M\}$.

We shall now introduce the notion of a function of two variables.

If every point M from a certain set $\{M\}$ of points of a Euclidean plane is associated with some number u in accordance with the known law, then a function $u = u(M)$ or $u = f(M)$ is said to be specified on the set $\{M\}$.

It should be pointed out that the concept of a function of two variables differs from the concept of a function of one variable formulated above only by the term "Euclidean plane" used instead of "Euclidean line". The concept of a function of three variables can be introduced by analogy, with the only difference that a set $\{M\}$ of points of a Euclidean space is taken instead of a set $\{M\}$ of points of a Euclidean plane.

Since point M of the Euclidean plane is defined by two coordinates x and y, and point M of the Euclidean space is defined by three coordinates x, y and z, the functions of two and three variables will be designated by $u = f(x, y)$ and $u = f(x, y, z)$ respectively. If the function $u = f(M)$ is defined on a set $\{M\}$, this set is called the *domain of definition of the function* $u = f(M)$. The number u corresponding to the given point M from the set $\{M\}$ will be called a *particular value of the function at point* M.

The totality $\{u\}$ of all the particular values of the function $u = f(M)$ is called the *set of values of that function.*

The notion of a graph can be introduced for a function of two variables, namely, *the graph of the function $u = f(x, y)$ is a surface whose points have the coordinates $(x, y, f(x, y))$.*

Let us consider some examples of functions of two and three variables.

1°. $u = \sqrt{4 - x^2 - y^2}$. The domain of definition of this function is a circle of radius 2 with centre at the origin, and the set of values is a closed interval $0 \leqslant u \leqslant 2$.

2°. $u = \dfrac{1}{\sqrt{x^2+y^2-4}}$. The domain of definition of this function is the set of points lying outside the circle of radius 2 with centre at the origin, and the set of values is an open half-line $u > 0$.

3°. $u = \sqrt{\cos(x^2 + y^2)}$. The domain of definition of this function is the set $\{M\}$ of points whose coordinates satisfy the inequality $\cos(x^2 + y^2) \geqslant 0$. This inequality is equivalent to the inequalities $0 \leqslant x^2 + y^2 \leqslant \pi/2$, $2k\pi - \pi/2 \leqslant x^2 + y^2 \leqslant 2k\pi + \pi/2$, $k = 1, 2, \ldots$ Thus $\{M\}$ consists of a circle of radius $\sqrt{\pi/2}$ with centre at point O (0, 0) and ring-like regions (Fig. 14.1).

4°. $u = \ln xyz$. The domain of definition of this function is the set $\{M\}$ of points whose coordinates satisfy the inequality $xyz > 0$, and the set of values is the entire number line $-\infty < u < +\infty$.

5°. $u = x^2 + y^2 + z^2$. The domain of definition of this function is the entire Euclidean space, and the set of values is the half-line $u \geqslant 0$.

14.1.4. The concepts of the m-dimensional coordinate space and m-dimensional Euclidean space. *The set of various ordered totalities* $(x_1, x_2, \ldots, x_m)$ *of the* m *numbers* $x_1, x_2, \ldots, x_m$ *is termed the* **m-dimensional coordinate space** A^m.

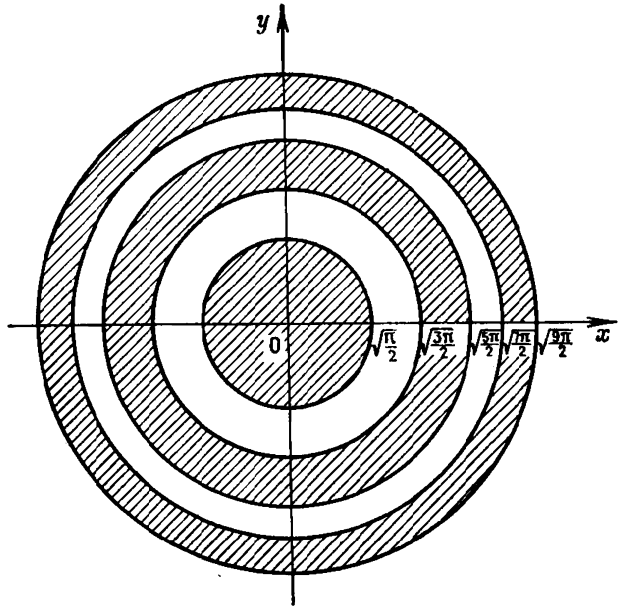


Fig. 14.1

In this case every ordered totality $(x_1, x_2, \ldots, x_m)$ will be called a point of that space and denoted by a single letter M. The numbers $x_1, x_2, \ldots, x_m$ are called the coordinates of point M. The notation $M(x_1, x_2, \ldots, x_m)$ means that point M has the coordinates $x_1, x_2, \ldots, x_m$. Now we shall introduce the notion of the m-dimensional Euclidean space. *The coordinate space* A^m *is called the* **m-dimensional Euclidean space** E^m *if the distance** $\rho(M', M'')$ *between any*

* The Euclidean m-dimensional space is the so-called *metric space*. The arbitrary set $\{M\}$ whose elements are called points is termed a metric space if there exists a rule according to which any two points M' and M'' of the set $\{M\}$ are associated with a certain number $\rho(M', M'')$ called the *distance* between the two points. The rule must make it possible to satisfy the following axioms (*axioms of a metric space*): (1) $\rho(M', M'') = \rho(M'', M')$ for any M' and M'' (the symmetry of the distance); (2) $\rho(M', M'') \geqslant 0$, for any M' and M''

two points $M'(x_1', x_2', \ldots, x_m')$ *and* $M''(x_1'', x_2'', \ldots, x_m'')$ *of the coordinate space* A^m *is determined by the formula*

$$\rho(M', M'') = \sqrt{(x_1''-x_1')^2+(x_2''-x_2')^2+\ldots+(x_m''-x_m')^2}. \quad (14.1)$$

The concepts of the m-dimensional coordinate space A^m and the m-dimensional Euclidean space E^m are generalizations of the concepts of the coordinate space and the Euclidean space presented above.

14.1.5. The sets of points of the m-dimensional Euclidean space E^m. The symbol $\{M\}$ will be used to designate a certain set of points of the m-dimensional Euclidean space E^m. Let us consider some examples of sets in the m-dimensional Euclidean space E^m.

1°. The set $\{M\}$ of various points whose coordinates $x_1, x_2, \ldots, x_m$ satisfy the inequality $(x_1 - x_1^0)^2 + (x_2 - x_2^0)^2 + \ldots + (x_m - x_m^0)^2 \leqslant R^2$ is called the *m-dimensional sphere* of radius R with centre at point $M_0(x_1^0, x_2^0, \ldots, x_m^0)$. Thus, the m-dimensional sphere is defined as a set $\{M\}$ of various points, the distance ρ between each of which and some point M_0 (the centre of the sphere) satisfies the inequality $\rho(M, M_0) \leqslant R$. If the distance $\rho(M, M_0)$ between each point of the set $\{M\}$ and point M_0 complies with the strict inequality $\rho(M, M_0) < R$, then the set $\{M\}$ is called an *open m-dimensional sphere.*

2°. The set $\{M\}$ of points the distance between each of which and point M_0 complies with the relation $\rho(M, M_0) = R$ is called the *m-dimensional sphere* of radius R and centre at point M_0.

---

and if $\rho(M', M'') = 0$ then the points M' and M'' coincide; (3) for any three points M', M'' and M''' the inequality $\rho(M', M''') \leqslant \rho(M', M'') + \rho(M'', M''')$ is satisfied (the inequality of a triangle).

Let us verify that the m-dimensional Euclidean space we have introduced is really a metric space. Indeed, the validity of the first two axioms of a metric space is evident (see formula (14.1)). Let us verify the third axiom.

Assume that x_i', x_i'', x_i''' are the coordinates of the points M', M'' and M'''. We have $\rho^2(M', M''') = \sum_{i=1}^{m} [(x_i''' - x_i'') + (x_i'' - x_i')]^2 = \sum_{i=1}^{m} (x_i''' - x_i'')^2 + 2\sum_{i=1}^{m}(x_i''' - x_i'')(x_i'' - x_i') + \sum_{i=1}^{m} (x_i'' - x_i')^2$. Putting $x_i''' - x_i'' = a_i$ and $x_i'' - x_i' = b_i$ and using Bunyakovsky's inequality (see inequality (10.30) in Supplement 1 to Chap. 10), we find that $\sum_{i=1}^{m} (x_i''' - x_i'')(x_i'' - x_i') \leqslant \sqrt{\sum_{i=1}^{m} (x_i''' - x_i'')^2} \times \sqrt{\sum_{i=1}^{m} (x_i'' - x_i')^2}$. From this it follows that $\rho^2(M', M''') \leqslant \left(\sqrt{\sum_{i=1}^{m} (x_i''' - x_i'')^2} + \sqrt{\sum_{i=1}^{m} (x_i'' - x_i')^2}\right)^2$, i.e. $\rho(M', M''') \leqslant \rho(M', M'') + \rho(M'', M''')$.

3°. The set $\{M\}$ of points whose coordinates $x_1, x_2, \ldots, x_m$ satisfy the inequalities $|x_1 - x_1^0| \leqslant d_1$, $|x_2 - x_2^0| \leqslant d_2, \ldots, |x_m - x_m^0| \leqslant d_m$ is called an *m-dimensional coordinate parallelepiped.* In this case point $M_0(x_1^0, x_2^0, \ldots, x_m^0)$ is termed the centre of this m-dimensional parallelepiped. If the coordinates $x_1, x_2, \ldots, x_m$ of the points of the set $\{M\}$ satisfy the strict inequalities $|x_1 - x_1^0| < d_1$, $|x_2 - x_2^0| < d_2, \ldots, |x_m - x_m^0| < d_m$, then the set $\{M\}$ is termed an *open m-dimensional coordinate parallelepiped.*

Let us now introduce the concepts of the *ε-neighbourhood* of point M_0 of the Euclidean m-dimensional space and of the *rectangular neighbourhood* of that point M_0. *We shall use the term the* **ε-neighbourhood of point** $M_0(x_1^0, x_2^0, \ldots, x_m^0)$ *of the m-dimensional Euclidean space* E^m *for an open m-dimensional sphere of radius* ε *with centre at point* M_0. **The rectangular neighbourhood of point** $M_0(x_1^0, x_2^0, \ldots x_m^0)$ *of the m-dimensional Euclidean space is any open m-dimensional coordinate parallelepiped with centre at point* M_0.

The following obvious assertion holds true.

Any ε-neighbourhood of point M_0 *of the Euclidean m-dimensional space* E^m *contains a certain rectangular neighbourhood of that point. Any rectangular neighbourhood of point* M_0 *contains a certain ε-neighbourhood of point* M_0.

Assume that $\{M\}$ is some set of points of the m-dimensional Euclidean space E^m. We shall introduce the following notions.

Point M of the set $\{M\}$ *is called an* **interior** *point of the set if there exists a certain ε-neighbourhood of point M whose all points belong to* $\{M\}$.

Point M^* *is called a* **boundary** *point of the set* $\{M\}$ *if any ε-neighbourhood of that point contains both the points belonging to the set* $\{M\}$ *and the points not belonging to that set.*

The set $\{M\}$ *of the space* E^m *is called an* **open set** *or* **domain** *if any point of that set is an interior point.*

If every boundary point of the set $\{M\}$ *is a point of that set, then the set* $\{M\}$ *is called a* **closed** *set.*

If the set $\{M\}$ is a domain, then the set $\{\overline{M}\}$ obtained by adding to $\{M\}$ all the boundary points of that set is called a **closed** *domain.*

It should be pointed out that if all the points of the domain $\{M\}$ are in the interior of a certain sphere, then the domain is said to be **bounded.**

In what follows we shall need the notion of a *connected set*, so we shall first introduce the notion of a *continuous curve* in the multidimensional space E^m.

We shall use the term a continuous curve L in the space E^m *for the set* $\{M\}$ *of points of that space, whose coordinates* $x_1, x_2, \ldots, x_m$ *are*

---

* Note that in this case point M may not belong to the set $\{M\}$.

continuous functions of parameter t:

$$x_1 = \varphi_1(t),\ x_2 = \varphi_2(t),\ \ldots,\ x_m = \varphi_m(t),\ \alpha \leqslant t \leqslant \beta. \qquad (14.2)$$

We shall say that the points M' $(x_1', x_2', \ldots, x_m')$ and M'' $(x_1'', x_2'', \ldots, x_m'')$ of the space E^m *can be connected by means of a continuous curve* L if there exists a continuous curve L specified by parametric equations (14.2), such that

$$x_1' = \varphi_1(\alpha),\ x_2' = \varphi_2(\alpha),\ \ldots,\ x_m' = \varphi_m(\alpha),$$
$$x_1'' = \varphi_1(\beta),\ x_2'' = \varphi_2(\beta),\ \ldots,\ x_m'' = \varphi_m(\beta),$$

Here is the formulation of the notion of a connected set. *The set* $\{M\}$ *of the space* E^m *is said to be* **connected** *if its any two points can be connected by a continuous curve whose all points belong to that set.*

Remark. Note that sometimes the term domain is used to describe an open and connected set and not simply an open set.

Let us consider the following example.

The set $\{M\}$ of points of E^m specified by the equation

$$\frac{x_1^2}{a_1^2} + \frac{x_2^2}{a_2^2} + \ldots + \frac{x_m^2}{a_m^2} = 1 \qquad (14.3)$$

is called an m-dimensional *ellipsoid.* The points of an m-dimensional ellipsoid are the boundary points of the set $\{M\}$ of points M, whose coordinates satisfy the inequality

$$\frac{x_1^2}{a_1^2} + \frac{x_2^2}{a_2^2} + \ldots + \frac{x_m^2}{a_m^2} < 1.$$

This set is the set of interior points of an m-dimensional ellipsoid.

The reader will easily verify that the set of interior points of an m-dimensional ellipsoid is open and connected. It should be noted that an m-dimensional ellipsoid specified by relation (14.3) is a closed set.

The domain of definition of the function

$$u = \sqrt{\cos(x^2 + y^2)}$$

is an *unconnected* set (see Example 3° in 1.3 and Fig. 14.1).

Let us agree in conclusion to use the term the *neighbourhood of point* M for any open connected set containing M.

14.1.6. The concept of a function of *m* variables. We shall now introduce the notion of a function of m variables.

If every point M *from the set* $\{M\}$ *of points of an* m*-dimensional Euclidean space* E^m *is put into correspondence with some number* u *in accordance with a certain law, then a function* $u = u(M)$ *or* $u = f(M)$ *is said to be specified on the set* $\{M\}$, the set $\{M\}$ being called the *domain of definition of the function* $u = f(M)$.

The number u corresponding to the given point M from the set $\{M\}$ will be called a *particular value of the function at point M*. The totality $\{u\}$ of all the particular values of the function $u = f(M)$ is called the *set of values* of the given function. Since point M is defined by the coordinates $x_1, x_2, \ldots, x_m$, the function $u = f(M)$ of m variables can be also designated as $u = f(x_1, x_2, \ldots, x_m)$.

Let us consider some examples of functions of m variables.

1°. Assume $u = \sqrt{1 - x_1^2 - x_2^2 - \ldots - x_m^2}$. The domain of definition of this function is evidently an m-dimensional sphere of radius 1 with centre at point $O(0, 0, \ldots, 0)$. The set of values of the function being discussed is the closed interval $[0, 1]$.

2°. Assume $u = \dfrac{1}{\sqrt{1 - \dfrac{x_1^2}{a_1^2} - \dfrac{x_2^2}{a_2^2} - \ldots - \dfrac{x_m^2}{a_m^2}}}$. The domain of definition of the function is the set $\{M\}$ of the interior points of an m-dimensional ellipsoid. The set of values of this function is the half-line $u \geqslant 1$.

14.2. THE LIMITING VALUE OF A FUNCTION OF SEVERAL VARIABLES

14.2.1. Convergent sequences of points in m-dimensional Euclidean space E^m. Cauchy's test for convergence of a sequence. Let us consider the sequence of points $\{M_n\}$ in an m-dimensional Euclidean space E^m*. We shall formulate the following definition.

The sequence $\{M_n\}$ of points of the Euclidean space E^m is said to be **convergent** *if there is a point A such that for any positive number ε there can be indicated a number N** such that for $n \geqslant N$ there holds an inequality $\rho(M_n, A) < \varepsilon$. In this case point A is called the* **limit** *of the sequence $\{M_n\}$.* The following notation is used to designate the limit A of the sequence $\{M_n\}$:

$$\lim_{n \to \infty} M_n = A \text{ or } M_n \to A \text{ for } n \to \infty.$$

Let us prove the following lemma.

Lemma 1. *Assume that the sequence $\{M_n\}$ of points of the Euclidean space E^m converges to point A. Then the sequences $\{x_1^{(n)}\}, \{x_2^{(n)}\}, \ldots, \{x_m^{(n)}\}$ of the coordinates of points M_n converge to the corresponding coordinates $a_1, a_2, \ldots, a_m$ of point A and, conversely, if the sequences*

---

* The notion of a sequence of points in the Euclidean space E^m is defined as follows. Assume that every number n of the natural series of numbers $1, 2, \ldots, n, \ldots$ is associated with a point M_n from the Euclidean space E^m. The resulting series of points $M_1, M_2, \ldots, M_n, \ldots$ considered in the indicated order is called a *sequence* of points of the Euclidean space E^m. We shall use the abbreviated symbol $\{M_n\}$ to designate this sequence.

** Since the number N depends, in the general case, on ε, they sometimes write $N = N(\varepsilon)$.

$\{x_1^{(n)}\}$, $\{x_2^{(n)}\}$, . . ., $\{x_m^{(n)}\}$ of the coordinates of points M_n converge, respectively, to the numbers a_1, a_2, . . ., a_m, then the sequence $\{M_n\}$ converges to point A with coordinates a_1, a_2, . . ., a_m.

Proof. We shall prove the first part of the lemma. If the sequence $\{M_n\}$ converges to point A, then for any $\varepsilon > 0$ there is a number N such that for $n \geqslant N$ there holds an inequality $\rho(M_n, A) < \varepsilon$. Suppose $(x_1^{(n)}, x_2^{(n)}, \ldots, x_m^{(n)})$ are coordinates of point M_n and $(a_1, a_2, \ldots, a_m)$ are coordinates of point A. Then the inequality $\rho(M_n, A) < \varepsilon$ can be written in the form

$$\sqrt{(x_1^{(n)}-a_1)^2+(x_2^{(n)}-a_2)^2+\ldots+(x_m^{(n)}-a_m)^2}<\varepsilon, \tag{14.4}$$

whence it follows that for $n > N$ there hold inequalities

$$|x_1^{(n)}-a_1|<\varepsilon,\ |x_2^{(n)}-a_2|<\varepsilon,\ \ldots,\ |x_m^{(n)}-a_m|<\varepsilon.$$

In other words, the sequences $\{x_1^{(n)}\}$, $\{x_2^{(n)}\}$, . . ., $\{x_m^{(n)}\}$ of the coordinates of points M_n converge, respectively, to the numbers a_1, a_2, . . ., a_m. Let us now prove the converse assertion. We shall assume that the indicated sequences of the coordinates of points M_n converge, respectively, to the numbers a_1, a_2, . . ., a_m. Then, for any $\varepsilon > 0$ there are numbers $N_1, N_2, \ldots, N_m$ such that for $n \geqslant N_1$, $n \geqslant N_2$, . . ., $n \geqslant N_m$ there hold inequalities

$$|x_1^{(n)}-a_2|<\frac{\varepsilon}{\sqrt{m}},\ |x_2^{(n)}-a_2|<\frac{\varepsilon}{\sqrt{m}},\ \ldots,\ |x_m^{(n)}-a_m|<\frac{\varepsilon}{\sqrt{m}}$$

respectively. It follows that for $n \geqslant N = \max\{N_1, N_2, \ldots, N_m\}$ inequality (14.4) holds true. To put it otherwise, for $n \geqslant N$ there holds an inequality $\rho(M_n, A) \leqslant \varepsilon$, where A is a point of E^m with coordinates $a_1, a_2, \ldots, a_m$. Thus we see that the sequence $\{M_n\}$ converges to point A. The lemma has been proved.

Here is a formulation of the definition of the fundamental sequence of points in an m-dimensional Euclidean space. *The sequence $\{M_n\}$ of points of an m-dimensional Euclidean space is called* **fundamental** *or Cauchy's sequence if for any positive number ε there is a number N such that for $n \geqslant N$ and for any natural p there holds an inequality $\rho(M_{n+p}, M_n) < \varepsilon$.* The following test for convergence (Cauchy's test) is valid.

For the sequence $\{M_n\}$ of points of an m-dimensional Euclidean space to be convergent, it is necessary and sufficient for it to be fundamental.

To verify the validity of the formulated test, we must note that if the sequence $\{M_n\}$ is fundamental, the sequences $\{x_1^{(n)}\}$, $\{x_2^{(n)}\}$, . . ., $\{x_m^{(n)}\}$ of the coordinates of points M_n are fundamental too, and, conversely, if the indicated sequences of the coordinates are fundamental, then the sequence $\{M_n\}$ is also fundamental. Then we must apply Cauchy's test for number sequences to the sequences $\{M_n\}$ of coordinates of points and Lemma 1 from the present subsection.

14.2.2. Some properties of bounded sequences of points in an m-dimensional Euclidean space. We shall now introduce the concept of a bounded sequence of points in an m-dimensional Euclidean space. *The sequence $\{M_n\}$ of points of an m-dimensional Euclidean space is said to be* **bounded** *if there exists a number $a > 0$ such that for all n there holds an inequality $\rho(O, M_n) \leqslant a$, where O is a point whose all coordinates are equal to zero.* In other words, the sequence $\{M_n\}$ is bounded if all the points M_n of that sequence are in the interior or on the boundary of a certain sphere with centre at the origin.

The following *fundamental* theorem holds true.

Theorem 14.1 (Bolzano-Weierstrass's theorem). *A convergent subsequence can be isolated from any bounded sequence $\{M_n\}$ of points of an m-dimensional Euclidean space.*

Proof. Let us ascertain, first of all, that the sequences $\{x_1^{(n)}\}$, $\{x_2^{(n)}\}$, ..., $\{x_m^{(n)}\}$ of coordinates of points M_n are bounded. Indeed, since the sequence $\{M_n\}$ is bounded, the inequality $\rho(O, M_n) \leqslant a$ is satisfied for all n. Since $\rho(O, M_n) = \sqrt{x_1^{(n)2} + x_2^{(n)2} + \ldots + x_m^{(n)2}}$, it follows that for all n there hold inequalities $|x_1^{(n)}| \leqslant a$, $|x_2^{(n)}| \leqslant a$, ..., $|x_m^{(n)}| \leqslant a$. In other words, the sequences $\{x_1^{(n)}\}$, $\{x_2^{(n)}\}$, ..., $\{x_m^{(n)}\}$ of coordinates of points M_n are bounded. By virtue of Bolzano-Weierstrass's theorem for number sequences (see 3.4.4), a sequence $\{x_1^{(n_{k_1})}\}$ converging to some number a_1 can be isolated from the sequence $\{x_1^{(n)}\}$. Let us consider the subsequence $\{x_2^{(n_{k_1})}\}$ of the sequence of the second coordinates of points M_n. In accordance with the same theorem, a subsequence $\{x_2^{(n_{k_2})}\}$ converging to some number a_2 can be isolated from the subsequence $\{x_2^{(n_{k_1})}\}$. Note that the subsequence $\{x_1^{(n_{k_2})}\}$ of the sequence $\{x_1^{(n_{k_1})}\}$ converges to a number a_1. Thus, the subsequences $\{x_1^{(n_{k_2})}\}$ and $\{x_2^{(n_{k_2})}\}$ converge to the numbers a_1 and a_2 respectively. It is evident that if we isolate a subsequence $\{x_3^{(n_{k_3})}\}$ converging to some number a_3 from the subsequence $\{x_3^{(n_{k_2})}\}$ of the sequence of the third coordinates of points M_n, then the subsequences $\{x_1^{(n_{k_3})}\}$, $\{x_2^{(n_{k_3})}\}$, $\{x_3^{(n_{k_3})}\}$ converge to the numbers a_1, a_2 and a_3 respectively. Continuing the argument, we shall finally obtain a subsequence $\{x_m^{(n_{k_m})}\}$ converging to some number a_m of the sequence of the mth coordinates of points M_n, with the subsequences $\{x_1^{(n_{k_m})}\}$, $\{x_2^{(n_{k_m})}\}$, ..., $\{x_m^{(n_{k_m})}\}$ converging to the numbers a_1, a_2, ..., a_m respectively. But then, by virtue of Lemma 1, the subsequence $\{M_{n_{k_m}}\}$ of the sequence of points

$\{M_n\}$ converges to point A with coordinates $a_1, a_2, \ldots, a_m$. We have proved the theorem.

Remark. The limit A of the sequence $\{M_n\}$ of points belonging to the closed set $\{M\}$ belongs to that set too. To verify this fact, it is sufficient to note that in any ε-neighbourhood of point A there are points M_n, that is, points of the set $\{M\}$ and, therefore, point A is either an interior or a boundary point of $\{M\}$ and, consequently, belongs to $\{M\}$.

14.2.3. The concept of the limiting value of a function of several variables. Let us consider the function $u = f(M)$ specified on the set $\{M\}$ of an m-dimensional Euclidean space and point A of that set which may not belong to the set $\{M\}$ but which possesses the property such that in any ε-neighbourhood of that point there is at least one point of the set $\{M\}$ different from A.

Definition 1. *The number b is called the* **limiting value of the function** $u = f(M)$ **at point** A *(or the* **limit of the function** *at $M \to A$), if for any sequence $M_1, M_2, \ldots, M_n, \ldots$ of points of the set $\{M\}$, converging to A, whose terms M_n are different from A* ($M_n \neq A$), the corresponding sequence $f(M_1), f(M_2), \ldots, f(M_n), \ldots$ of the values of the function converges to b.*

This definition is known as the definition of the limiting value of a function by means of sequences. We can give another definition of the limiting value of a function with the use of "ε-δ"-terminology.

Definition 2. *The number b is called the limiting value of the function $u = f(M)$ at point A if for any positive number ε there is a positive number δ such that for all the points M belonging to the domain of definition of the function and satisfying the condition $0 < \rho(M, A) < \delta$, the inequality $|f(M) - b| < \varepsilon$ is complied with.*

Remark. Definitions 1 and 2 of the limiting value of a function are equivalent. The validity of this assertion can be proved in the same way as the equivalence of the two definitions of the limiting value of a function of one variable. The following notation is used for the limiting value b of the function $u = f(M)$ at point A:

$$\lim_{M \to A} f(M) = b, \text{ or } \lim_{\substack{x_1 \to a_1 \\ x_2 \to a_2 \\ \ldots\ldots \\ x_m \to a_m}} f(x_1, x_2, \ldots, x_m) = b,$$

where $a_1, a_2, \ldots, a_m$ are the coordinates of point A.

We shall now formulate the definition of the limiting value of a function in the case of point M tending to infinity.

Definition 3. *The number b is called the* **limiting value of the function** $u = f(M)$ *at $M \to \infty$ (or the* **limit of the function** *at $M \to \infty$)*

* This requirement is due, in particular, to the fact that the function $u = f(M)$ may not be defined at point A.

if for any positive number ε *there is a positive number* a *such that for all* M *belonging to the domain of definition of the function and complying with the condition* $\rho(O, M) > a$, *the inequality* $|f(M) - b| < \varepsilon$ *is satisfied.*

Arithmetic operations carried out on functions of m variables having a limiting value at point A lead to functions which also possess a limiting value at point A. In fact, the following assertion holds true.

Suppose the functions $f(M)$ *and* $g(M)$ *possess limiting values* b *and* c *at point* A. *Then the functions* $f(M)+g(M)$, $f(M)-g(M)$, $f(M)\cdot g(M)$ *and* $\frac{f(M)}{g(M)}$ *possess at point* A *limiting values (a quotient when* $c \neq 0$*) equal, respectively, to* $b+c$, $b-c$, $b\cdot c$, b/c.

This assertion can be proved by a complete analogy with Theorem 4.1.

14.2.4. Infinitesimal functions. *The function* $u = f(M)$ *is said to be infinitesimal at point* A *(at* $M \to A$*) if* $\lim\limits_{M \to A} f(M) = 0$.

It is easy to verify that the function $f(M) = (x_1 - a_1)^{n_1} + \ldots + (x_m - a_m)^{n_m}$, where $n_1, \ldots, n_m$ are positive numbers, is infinitesimal at point $A(a_1, a_2, \ldots, a_m)$*.

If the function $u = f(M)$ *possesses at point* A *a limiting value equal to* b, *then the function* $\alpha(M) = f(M) - b$ *is infinitesimal at point* A. Indeed, $\lim\limits_{M \to A} \alpha(M) = \lim\limits_{M \to A} (f(M) - b) = \lim\limits_{M \to A} f(M) - \lim\limits_{M \to A} b = 0$. Making use of this result we shall obtain a special representation for the function possessing at point A a limiting value equal to b:

$$f(M) = b + \alpha(M), \text{ where } \lim_{M \to A} \alpha(M) = 0.$$

A comparison between infinitesimal functions of several variables can be carried out by a complete analogy with that indicated in 4.2.3 for infinitesimal functions of one variable. It should be pointed out that by analogy with the case of one variable the symbol $o(\beta)$ will denote any infinitesimal, at point A, function of the higher order of smallness than the infinitesimal, at point A, function $\beta(M)$.

14.2.5. A necessary and sufficient condition of the existence of the limiting value of a function (Cauchy's test). We shall say that the function $f(M)$ satisfies *Cauchy's condition* at point $M = A$ if for any positive number ε there is a positive number δ such that whatever the two points M' and M'' from the domain of definition of the function $f(M)$ satisfying the inequalities $0 < \rho(M', A) < \delta$, $0 < \rho(M'', A) < \delta$, the following inequality holds for the respec-

* It is sufficient to take into account that every function of one variable $f(x_k) = (x_k - a_k)^{n_k}$ is infinitesimal at point $x_k = a_k$.

tive values of the functions:

$$| f(M') - f(M'') | < \varepsilon.$$

The following *fundamental* theorem holds true.

Theorem 14.2 (Cauchy's test). *For the function $f(M)$ to possess a finite limiting value at point $M = A$, it is necessary and sufficient for the function $f(M)$ to satisfy Cauchy's test at that point.* The proof of this theorem is completely analogous to the proof of Theorem 8.2 and can be obtained from it by replacing the letters x and a by the letters M and A using the symbol $\rho(M, A)$ instead of the expressions of the kind of $| x - a |$.

14.2.6. Repeated limiting values. For the function $u = f(x_1, x_2, \ldots, x_m)$ of several variables the notion of the limiting value can be defined from one of the variables x_k with the values of the rest of the variables being fixed. In this connection there appears the notion of the *repeated limiting value.* We shall elucidate this notion on the example of the function $u = f(x, y)$ of two variables x and y. Assume that the function $u = f(x, y)$ is specified in some rectangular neighbourhood $| x - x_0 | < d_1$, $| y - y_0 | < d_2$ of point $M_0(x_0, y_0)$, with the exception, perhaps, of the point M_0 itself. Suppose that for every fixed y satisfying the condition $0 < | y - y_0 | < d_2$ there is a limiting value of the function $u = f(x, y)$ of one variable x at point $x = x_0$:

$$\lim_{\substack{x \to x_0 \\ y \text{ is fixed}}} f_1(x, y) = \varphi(x),$$

and suppose, in addition, there exists a limiting value b of the function $\varphi(y)$ at point $y = y_0$:

$$\lim_{y \to y_0} \varphi(y) = b.$$

In this case we say that there exists a *repeated* limiting value b for the function $u = f(x, y)$ at point M_0 which is designated as follows:

$$\lim_{y \to y_0} \lim_{x \to x_0} f(x, y) = b.$$

The repeated limiting value

$$\lim_{x \to x_0} \lim_{y \to y_0} f(x, y)$$

is defined in a similar way.

Let us establish the sufficient conditions of equality of the two repeated limiting values we have introduced.

Theorem 14.3. *Assume that the function $u = f(x, y)$ is defined in some rectangular neighbourhood $| x - x_0 | < d_1$, $| y - y_0 | < d_2$ of the point $M_0(x_0, y_0)$ and possesses at this point a limiting value equal to b. Assume, in addition, that for any fixed x, $0 < | x - x_0 | < d_1$, there exists a limiting value $\psi(x) = \lim_{y \to y_0} f(x, y)$ and for any fixed y, $0 < | y - y_0 | < d_2$, there exists a limiting value $\varphi(y) = \lim_{x \to x_0} f(x, y)$. Then the repeated limiting values $\lim_{x \to x_0} \lim_{y \to y_0} f(x, y)$ and $\lim_{y \to y_0} \lim_{x \to x_0} f(x, y)$ exist and are equal to b.*

Proof. Since the function $u = f(x, y)$ possesses a limiting value b at point $M_0(x_0, y_0)$, there is a $\delta > 0$ for any $\varepsilon > 0$, such that the inequality $| f(x, y) - b | < \varepsilon$ holds at $| x - x_0 | < \delta$ and $| y - y_0 | < \delta$. Thus, in the

rectangular neighbourhood $|x - x_0| < \delta$ and $|y - y_0| < \delta$ of point M_0 the values of the function $f(x, y)$ differ from b by not more than ε. But then for x and y satisfying the inequalities $|x - x_0| < \delta$ and $|y - y_0| < \delta$, the limiting values $\psi(x)$ and $\varphi(y)$ indicated in the formulation of the theorem also differ from b by not more than ε. Consequently, the limiting values of these functions at points x_0 and y_0 respectively exist and are equal to b. We have proved the theorem.

The notion of the repeated limit can be defined for the so-called double sequences $\{a_{mn}\}$ whose terms a_{mn} are specified by two indices, m and n. Namely, the symbol $\lim\limits_{n\to\infty} \lim\limits_{m\to\infty} a_{mn}$ signifies that first we determine the sequence $\{b_n\}$, $b_n = \lim\limits_{m\to\infty} a_{mn}$ and then find the limit of that sequence $\{b_n\}$.

By way of example, let us consider the double sequence $\{a_{mn}\}$, where $a_{mn} = \cos^m 2\pi n! x$, x being a fixed number. Let us prove that

$$\lim_{n\to\infty} \lim_{m\to\infty} \cos^m 2\pi n! x = \begin{cases} 1, & \text{if } x \text{ is a rational number,} \\ 0, & \text{if } x \text{ is an irrational number.} \end{cases}$$

Indeed, if $x = \frac{p}{q}$, where p and q are integers, then for $n \geqslant q$ we have $\cos 2\pi n! x = 1$ and, therefore, $\lim\limits_{\substack{m\to\infty \\ n\geqslant q}} \cos^m 2\pi n! x = 1$. In other words, if x is a rational number, then $\lim\limits_{n\to\infty} \lim\limits_{m\to\infty} \cos^m 2\pi n! x = 1$. Now if x is an irrational number, then inequality $|\cos 2\pi n! x| < 1$ is valid for any n and, therefore, $\lim\limits_{m\to\infty} \cos^m 2\pi n! x = 0$, i.e. $\lim\limits_{n\to\infty} \lim\limits_{m\to\infty} \cos^m 2\pi n! x = 0$.

Remark. Making use of the result we have obtained, we can specify Dirichlet's function analytically (see 4.1.1) as a repeated limit $\lim\limits_{n\to\infty} \lim\limits_{m\to\infty} \cos^m 2\pi n! x$.

14.3. CONTINUOUS FUNCTIONS OF SEVERAL VARIABLES

14.3.1. The definition of the continuity of a function of several variables. Assume that point A belongs to the domain of definition of the function $u = f(M)$ of several variables and any ε-neighbourhood of point A contains points from the domain of definition of that function different from A.

Definition 1. *The function $u = f(M)$ is said to be* **continuous at point** *A if the limiting value of that function at point A exists and is equal to the particular value $f(A)$.* Note that since $A = \lim\limits_{M\to A} M$, the condition of continuity of the function can be written in the following form:

$$\lim_{M\to A} f(M) = f\left(\lim_{M\to A} M\right).$$

The points at which the function does not possess the continuity property are called the *points of discontinuity* of that function.

Let us formulate the difinition of the continuity of a function, using the definition of a limiting value of a function in terms of ε and δ.

Definition 2. *The function $u = f(M)$ is said to be continuous at point A if for any positive number ε there is a positive number δ such*

that for all the points M from the domain of definition of the function satisfying the condition $\rho(M, A) < \delta$, there holds an inequality $|f(M) - f(A)| < \varepsilon$.

Definition 3. *The function $u = f(M)$ is said to be* **continuous on the set** $\{M\}$ *if it is continuous at every point of that set.*

We shall use the term *increment or total increment of the function $u = f(M)$ at point A for the function Δu specified by the formula*

$$\Delta u = f(M) - f(A), \tag{14.5}$$

where M is any point from the domain of definition of the function. Suppose points A and M have the respective coordinates $a_1, a_2, \ldots, a_m$ and $x_1, x_2, \ldots, x_m$. Making use of the notation $x_1 - a_1 = \Delta x_1$, $x_2 - a_2 = \Delta x_2, \ldots, x_m - a_m = \Delta x_m$, we shall obtain the following expression for the increment of the function Δu corresponding to the increments of the arguments $\Delta x_1, \ldots, \Delta x_m$:

$$\Delta u = f(a_1 + \Delta x_1, a_2 + \Delta x_2, \ldots, a_m + \Delta x_m) - \\ - f(a_1, a_2, \ldots, a_m). \tag{14.6}$$

It is evident that *for the function $u = f(M)$ to be continuous at point A, it is necessary and sufficient that its increment Δu be an infinitesimal function at point A*, i.e. it is necessary and sufficient that

$$\lim_{M \to A} \Delta u = \lim_{M \to A} (f(M) - f(A)) = 0 \text{ or } \lim_{\substack{\Delta x_1 \to 0, \\ \Delta x_2 \to 0, \\ \ldots\ldots \\ \Delta x_m \to 0}} \Delta u = 0. \tag{14.7}$$

Condition (14.7) will be called *the difference form of the condition of continuity of the function $u = f(M)$ at point A.*

For the function $u = f(x_1, x_2, \ldots, x_m)$ of several variables the concept of continuity can be defined with respect to one of the variables, the values of the other variables being fixed. To define this concept, let us consider the so-called *partial increments* of the function $u = f(x_1, x_2, \ldots, x_m)$ at point $M(x_1, x_2, \ldots, x_m)$ belonging to the domain of definition of the function. Let us fix all the arguments except for the first and impart to the first argument an arbitrary increment Δx_1 such that the point with the coordinates $x_1 + \Delta x_1, x_2, \ldots, x_m$ is in the domain of definition of the function. The corresponding increment of the function is called a *partial increment** of the function at point $M(x_1, x_2, \ldots, x_m)$ corresponding to the increment Δx_1 of the argument x_1 and is denoted as $\Delta_{x_1} u$. Thus we have

$$\Delta_{x_1} u = f(x_1 + \Delta x_1, x_2, \ldots, x_m) - f(x_1, x_2, \ldots, x_m). \tag{14.8}$$

* The term "partial increment" is used to distinguish this increment from the total increment (14.6) corresponding to the arbitrary increments $\Delta x_1, \Delta x_2, \ldots, \Delta x_m$ of all the arguments $x_1, x_2, \ldots, x_m$.

A similar procedure is used to determine the partial increments of the function corresponding to the increments of other arguments:

$$\left.\begin{aligned} &\Delta_{x_2} = f(x_1,\ x_2 + \Delta x_2, x_3,\ \ldots,\ x_m) - f(x_1,\ x_2,\ \ldots,\ x_m), \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &\Delta_{x_m} u = f(x_1,\ x_2,\ \ldots,\ x_{m-1},\ x_m + \Delta x_m) - f(x_1,\ x_2,\ \ldots\ x_m). \end{aligned}\right\} \quad (14.9)$$

Let us now introduce the concept of continuity of the function $u = f(x_1, x_2, \ldots, x_m)$ with respect to one of the variables.

The function $u = f(x_1, x_2, \ldots, x_m)$ is said to be continuous at point $M(x_1, x_2, \ldots, x_m)$ with respect to the variable x_k if the partial increment $\Delta_{x_k} u$ of that function at point M is an infinitesimal function of Δx_k, that is, if

$$\lim_{\Delta x_k \to 0} \Delta_{x_k} u = 0. \quad (14.10)$$

If the values of all the variables, except for the variable x_k, are fixed, then the function $u = f(x_1, x_2, \ldots, x_m)$ is a function of that one variable. Note that the continuity of a function with respect to the variable x_k implies the continuity of the indicated function of one variable. It is evident that from the condition of continuity of the function $u = f(x_1, x_2, \ldots, x_m)$ at the given point M follows the continuity of that function at point M with respect to each of the variables $x_1, x_2, \ldots, x_m$. However, the continuity of the function at point M with respect to each of the variables $x_1, x_2, \ldots, x_m$ does not imply, in general, the continuity of the function at that point. To verify this fact, let us consider the following examples.

1°. We shall say that the function $u = f(M) = f(x, y)$ is continuous at point M on a certain straight line passing through that point if for any sequence of points $\{M_n\}$ of that line converging to point M the corresponding sequence $\{f(M_n)\}$ of the values of the function has as its limit the partial value $f(M)$ of the function at point M. Since on the straight line the function $u = f(x, y)$ is a function of one variable, the notion of the continuity of a function on a line evidently coincides with the notion of the continuity of the indicated function of one variable. In particular, the continuity of the function at point M with respect to the separate variables x and y is its continuity on straight lines which pass through point M and are parallel to the coordinate axes. Let us prove that the function

$$u = \begin{cases} \dfrac{xy}{x^2+y^2} & \text{for } x^2+y^2 \neq 0, \\ 0 & \text{for } x^2+y^2 = 0 \end{cases}$$

is continuous at point $O(0, 0)$ with respect to each variable x and y, that is, continuous on each coordinate axis, but not continuous on all the rest straight lines passing through that point and, therefore,

not continuous at point O. Every straight line different from the coordinate axes and passing through point O (0, 0) can be represented by the equation $y = kx$, where $k \neq 0$. It is evident that on such a straight line all the values of the function are constant and are equal to $k/(1 + k^2)$. Consequently, if the sequence $\{M_n\}$ of points of such a line, different from O, converges to point O, then the corresponding sequence of the values of the function has the limit $k/(1 + k^2)$. Since at $k \neq 0$ this limit is nonzero and does not coincide with the partial value of the function at point O, the function is discontinuous at that point of the line being considered. The continuity of the function on the coordinate axes follows from the fact that its values on those axes are equal to zero.

One can think that if a function of two variables is continuous on any line passing through that point, then the function is continuous at the indicated point. The following example shows that in the general case it is not so.

2°. Let us consider the function

$$u = f(M) = \begin{cases} \dfrac{x^2 y}{x^4 + y^2} & \text{for } x^4 + y^2 \neq 0, \\ 0 & \text{for } x^4 + y^2 = 0. \end{cases}$$

We shall prove that although the indicated function is continuous on any line passing through the point O (0, 0) it is not continuous at that point. Indeed, the values of the function on the line $y = kx$ are equal to $\dfrac{kx}{x^2 + k^2}$ and, therefore, $u \to 0$ as $x \to 0$. The continuity of this function on the y-axis follows from the fact that its values on that axis are equal to zero. On the other hand, the values of the function on the parabola $y = px^2$ are constant and equal to $\dfrac{p}{1 + p^2}$ and, therefore, the limiting value of the function is also equal to $\dfrac{p}{1 + p^2}$ when point M tends to point O on the indicated parabola. Since at $p \neq 0$ this limit is nonzero and does not coincide with the particular value of the function at point O, the function is discontinuous at that point.

14.3.2. Main properties of discontinuous functions of several variables. Here we shall enumerate the main properties of continuous functions of several variables. Since in the main the proof of these properties is similar to that of the corresponding properties of functions of one variable, we shall restrict ourselves, as a rule, to brief explanations, leaving the details of the proof to the reader.

1°. **Arithmetic operations on continuous functions.** The following assertion holds true.

Assume that the functions $f(M)$ and $g(M)$ are continuous at point A. Then the functions $f(M) + g(M)$, $f(M) - g(M)$, $f(M) \times g(M)$ and $\dfrac{f(M)}{g(M)}$ are continuous at point A (a quotient if $g(A) \neq 0$). The proof of this assertion is quite analogous to the proof of Theorem 4.2.

2°. **Continuity of a composite function.** We shall now introduce the concept of a composite function of several variables. Suppose that the functions

$$\left.\begin{aligned} x_1 &= \varphi_1(t_1, t_2, \ldots, t_k), \\ x_2 &= \varphi_2(t_1, t_2, \ldots, t_k), \\ &\ldots\ldots\ldots\ldots \\ x_m &= \varphi_m(t_1, t_2, \ldots, t_k) \end{aligned}\right\} \tag{14.11}$$

are defined on the set $\{N\}$ of the Euclidean space E^k ($t_1, t_2, \ldots, t_k$ being the coordinates of the points in that space). Then each point $N(t_1, t_2, \ldots, t_k)$ from the set $\{N\}$ is put into correspondence with a point $M(x_1, x_2, \ldots, x_m)$ of the Euclidean space E^m with the aid of formulas (14.11). Let us denote by $\{M\}$ the set of all such points. Assume that $u = f(x_1, x_2, \ldots, x_m)$ is a function of m variables defined on the set $\{M\}$. In that case we shall say that defined on the set $\{N\}$ of the Euclidean space E^k is a *composite function* $u = f(x_1, x_2, \ldots, x_m)$, where $x_1, x_2, \ldots, x_m$ are the functions of the variables $t_1, t_2, \ldots, t_k$, the functions being specified by relations (14.11). The following assertion holds true.

Assume that the functions $x_1 = \varphi_1(t_1, t_2, \ldots, t_k)$, $x_2 = \varphi_2(t_1, t_2, \ldots, t_k)$, $\ldots$, $x_m = \varphi_m(t_1, t_2, \ldots t_k)$ *are continuous at point* $A(a_1, a_2, \ldots, a_k)$, *and the function* $u = f(x_1, x_2, \ldots, x_m)$ *is continuous at point* $B(b_1, b_2, \ldots, b_m)$, *where* $b_i = \varphi_i(a_i, a_2, \ldots, a_k)$, $i = 1, 2, \ldots, m$. *Then the composite function* $u = f(x_1, x_2, \ldots, x_m)$, *where* $x_1, x_2, \ldots, x_m$ *are the functions of the arguments* $t_1, t_2, \ldots, t_k$ *specified above, is continuous at point* $A(a_1, a_2, \ldots, a_k)$. We shall carry out the proof of this assertion by stages. Suppose that $\{N_n\}$, $N_n \neq A$, is an arbitrary, converging to A, sequence of points from the domain $\{N\}$ of definition of the functions $\varphi_i(t_1, t_2, \ldots, t_k)$ and $\{M_n\}$ is the corresponding sequence of points whose coordinates $x_i^{(n)}$ are equal to $\varphi_i(t_1^{(n)}, t_2^{(n)}, \ldots, t_k^{(n)})$. By virtue of continuity of the functions φ_i at point A, the sequence $\{M_n\}$ converges to point $B(b_1, b_2, \ldots, b_m)$ (the case is not excluded where points M_n coincide with point B). The function $u = f(x_1, x_2, \ldots, x_m)$ being continuous at point B, the sequence $\{f(M_n)\}$ converges to $f(B)$. But this sequence is the sequence of values of the composite function corresponding to the sequence $\{N_n\}$, converging to A, of points of the domain of its definition. Since we have ascertained that the sequence $\{f(M_n)\}$ converges to the particular value $f(B)$, we have proved the continuity of a composite function.

Remark. The proof we have carried out is a generalization of the proof of Theorem 4.5 on the continuity of a composite function of one variable to the case of several variables.

3°. **The theorem on the stability of sign of a continuous function.**

Theorem 14.4. *If the function $u=f(M)$ is continuous at point A of the Euclidean space E^m and if $f(A) \neq 0$, then there exists a δ-neighbourhood of point A within which, at all the points of its definition, the function $f(M)$ does not vanish and is of the sign coinciding with the sign of $f(M)$.* The validity of this theorem follows immediately from the definition of continuity of a function in the "ε-δ" terminology.

4°. **Theorem on the passage of a continuous function through any intermediate value.**

Theorem 14.5. *Assume that the function $u=f(M)$ is continuous at all the points of the connected set $\{M\}$ of the Euclidean space E^m, with $f(A)$ and $f(B)$ being the values of that function at points A and B of $\{M\}$. Assume, furthermore, that C is any number contained between $f(A)$ and $f(B)$. Then, on any continuous curve L connecting points A and B and lying entirely in $\{M\}$ there is a point N such that $f(N)=C$.*

Proof. Assume that

$$x_1 = \varphi_1(t),\ x_2 = \varphi_2(t),\ \ldots,\ x_m = \varphi_m(t),\ \alpha \leqslant t \leqslant \beta$$

are equations of the continuous curve L connecting points A and B of the set $\{M\}$ and lying entirely in $\{M\}$ (see 1.5). Defined on the closed interval $[\alpha, \beta]$ is a composite function $u=f(x_1, x_2, \ldots, x_m)$, where $x_i=\varphi_i(t)$, $i=1, 2, \ldots, m$, $\alpha \leqslant t \leqslant \beta$. The values of this function on the closed interval $[\alpha, \beta]$ evidently coincide with the values of the function $u=f(M)$ on the curve L. The indicated composite function of one variable t is continuous on the interval $[\alpha, \beta]$ by virtue of the assertion made in 2° of the present subsection, and, in accordance with Theorem 8.6, assumes a value C at some point ξ of the interval $[\alpha, \beta]$. Therefore, the equality $f(N)=C$ is valid at point N of the curve L with the coordinates $\varphi_1(\xi)$, $\varphi_2(\xi)$, ..., $\varphi_m(\xi)$. We have proved the theorem.

5°. **The boundedness of a function continuous on a closed bounded set.**

Theorem 14.6 (Weierstrass's first theorem). *If the function $u=f(M)$ is continuous on a closed bounded set $\{M\}$, it is bounded on that set.* We shall discuss the proof of the boundedness of $u=f(M)$ from above. We shall suppose that $u=f(M)$ is not bounded from above on $\{M\}$. Let us isolate (as in the proof of the analogous Theorem 8.7) a sequence $\{M_n\}$ of points of the set $\{M\}$ for which $f(M_n) > n$. By virtue of Bolzano-Weierstrass's theorem (see 2.2) we can isolate from $\{M_n\}$ a convergent subsequence $\{M_{k_n}\}$ whose limit M, in accordance with the remark to Bolzano-Weierstrass's theorem, belongs to the set $\{M\}$. The sequence $\{f(M_{k_n})\}$ is evidently infinitely large. On the other hand, by virtue of the continuity of the function at

point M, this sequence $\{f(M_{k_n})\}$ must converge to $f(M)$. The contradiction we have obtained proves the theorem.

6°. **The attainment of the supremum and infimum by the function continuous on a closed bounded set.**

Theorem 14.7 (Weierstrass's second theorem). *If the function $u = f(M)$ is continuous on the closed bounded set $\{M\}$, it attains its supremum and infimum on that set.* The proof of this theorem is quite analogous to the proof of Theorem 8.8 (Weierstrass's second theorem for a function of one variable).

7°. **The concept of uniform continuity of a function of several variables.**. *The function $u = f(M)$ is said to be* **uniformly continuous** *on the set $\{M\}$* of the Euclidean space E^m if for any positive number ε there is a positive δ dependent solely on ε and such that for any two points M' and M'' of the set $\{M\}$ satisfying the condition $\rho(M', M'') < \delta$ there holds an inequality $|f(M'') - f(M')| < \varepsilon$.* The following theorem holds true.

Theorem 14.8 (theorem on uniform continuity). *A function continuous on the closed bounded set $\{M\}$ is uniformly continuous on that set.* The proof of this theorem is quite analogous to the proof of Theorem 10.2 and can be obtained from it by replacing the term "closed interval $[a, b]$" by the term "the set $\{M\}$", the letter x by the letter M and the expression of the kind $|x'' - x'|$ by the symbol $\rho(M', M'')$.

Remark. We shall use the term the *diameter* of the bounded set $\{M\}$ for the supremum of the numbers $\rho(M', M'')$, where M' and M'' are various points of the set $\{M\}$. Making use of the notion of the diameter of the set, let us note the following property of continuous functions on closed bounded sets. *Assume that the function $u = f(M)$ is continuous on the closed bounded set $\{M\}$. Then, for any positive number ε there is a $\delta > 0$ such that on any closed subset $\{\overline{M}\}$ belonging to the set $\{M\}$, whose diameter is smaller than δ, the oscillation ω** of the function $f(M)$ is less than ε.* The proof of this property is quite analogous to the proof of the corollary of Theorem 10.2.

14.4. DERIVATIVES AND DIFFERENTIALS OF A FUNCTION OF SEVERAL VARIABLES

14.4.1. Partial derivatives of a function of several variables. Assume point $M(x_1, x_2, \ldots, x_m)$ to be an interior point of the domain of definition of the function $u = f(x_1, x_2, \ldots, x_m)$. Let us consider, at a given fixed point $M(x_1, x_2, \ldots, x_m)$, the ratio between the partial increment $\Delta_{x_k} u$ (see 14.3.1, formulas (14.8) and (14.9)) and the cor-

* It is assumed here that the set $\{M\}$ is dense in itself, i.e. in any ε-neighbourhood of every point M of the set there are points of the set $\{M\}$ different from M.

** The oscillation ω of the function $f(M)$ on the set $\{\overline{M}\}$ is the difference between the supremum and the infimum of the function $f(M)$ on that set.

responding increment Δx_k of the argument x_k:

$$\frac{\Delta_{x_k} u}{\Delta x_k} = \frac{f(x_1, x_2, \ldots, x_{k-1}, x_k + \Delta x_k, x_{k+1}, \ldots, x_m) - f(x_1, x_2, \ldots, x_m)}{\Delta x_k}. \tag{14.12}$$

Relation (14.12) is a function of Δx_k determined for all nonzero values of Δx_k for which point $M(x_1, x_2, \ldots, x_{k-1}, x_k + \Delta x_k, x_{k+1}, \ldots, x_m)$ belongs to the domain of definition of the function u.

Definition. *If there exists a limit of the ratio* (14.12) *of the partial increment* Δ_{x_k} *and the function at point* $M(x_1, x_2, \ldots, x_m)$ *to the corresponding increment* Δx_k *of the argument* x_k *as* $\Delta x_k \to 0$, *then this limit is called a* **partial derivative** *of the function* $u = f(x_1, x_2, \ldots, x_m)$ *at point* M *with respect to the argument* x_k *and is designated by one of the following symbols*:

$$\frac{\partial u}{\partial x_k}, \quad \frac{\partial f}{\partial x_k}, \quad u'_{x_k}, \quad f'_{x_k}.$$

It follows that

$$\frac{\partial u}{\partial x_k} = \lim_{\Delta x_k \to 0} \frac{\Delta_{x_k} u}{\Delta x_k}. \tag{14.13}$$

Note that the partial derivative of the function $u = f(x_1, x_2, \ldots, x_m)$ with respect to the argument x_k is an ordinary derivative of the function x_k of one variable, the values of the rest of the variables being fixed. Therefore, partial derivatives can be calculated in accordance with the rules governing the calculation of the derivatives of functions of one variable.

Examples.

1°. $u = \arctan \frac{x}{y}$, $\frac{\partial u}{\partial x} = \frac{y}{x^2 + y^2}$, $\frac{\partial u}{\partial y} = \frac{-x}{x^2 + y^2}$.

2°. $u = xe^{yz} + \ln(x - y + z)$, $\frac{\partial u}{\partial x} = e^{yz} + \frac{1}{x - y + z}$.

$$\frac{\partial u}{\partial y} = xze^{yz} - \frac{1}{x - y - z}, \quad \frac{\partial u}{\partial z} = xye^{yz} + \frac{1}{x - y + z},$$

3°. $u = \tan \sqrt{x^2 - yz}$, $\frac{\partial y}{\partial x} = \frac{x}{\sqrt{x^2 - yz} \cos^2 \sqrt{x^2 - yz}}$,

$$\frac{\partial u}{\partial y} = -\frac{z}{2\sqrt{x^2 - yz} \cos^2 \sqrt{x^2 - yz}}, \quad \frac{\partial u}{\partial z} = -\frac{y}{2\sqrt{x^2 - yz} \cos^2 \sqrt{x^2 - yz}}.$$

Remark 1. The existence of all partial derivatives of a function at a given point does not imply, in general, the continuity of the

function at that point. We have verified that the function

$$u = \begin{cases} \dfrac{xy}{x^2+y^2} & \text{for } x^2+y^2 \neq 0, \\ 0 & \text{for } x^2+y^2 = 0 \end{cases}$$

is not continuous at point $O\,(0, 0)$ (see Example 1° in 14.3.1). However, the indicated function possesses at that point partial derivatives with respect to x and y. This follows from the fact that $f(x, 0) \equiv 0$ and $f(0, y) \equiv 0$ and, therefore,

$$\left.\frac{\partial f}{\partial x}\right|_{(0,\,0)} = 0 \text{ and } \left.\frac{\partial f}{\partial y}\right|_{(0,\,0)} = 0.$$

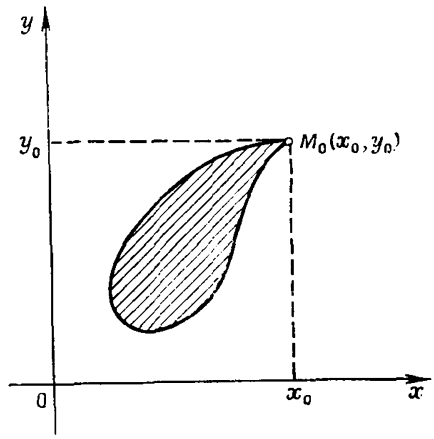


Fig. 14.2

Remark 2. We have defined the concept of partial derivatives for interior points of the domain of definition of a function. The definition of partial derivatives we have given does not apply, in general, to boundary points of the domain of definition of a function. In particular, it is due to the fact that at the boundary points of the domain of definition of a function it is not always possible to calculate partial increments of that function (this is the case, for instance, with the boundary point M_0 of the domain shown in Fig. 14.2). For that matter, at the boundary points of the domain of definition of a function partial derivatives are usually defined as the limiting values of those derivatives.

14.4.2. The concept of differentiability of a function of several variables. Let us recall that the increment (or total increment) of the function $u = f(x_1, x_2, \ldots, x_m)$ at point $M(x_1, x_2, \ldots, x_m)$, corresponding to the increments $\Delta x_1, \Delta x_2, \ldots, \Delta x_m$ of the arguments is the expression

$$\Delta u = f(x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots, x_m + \Delta x_m) - f(x_1, x_2, \ldots, x_m).$$

Definition. The function $u = f(x_1, x_2, \ldots, x_m)$ is said to be **differentiable** *at the given point $M(x_1, x_2, \ldots, x_m)$ if its total increment at that point can be represented in the form*

$$\Delta u = A_1\Delta x_1 + A_2\Delta x_2 + \ldots + A_m\Delta x_m + \alpha_1\Delta x_1 + \alpha_2\Delta x_2 + \\ + \ldots + \alpha_m\Delta x_m, \qquad (14.14)$$

where $A_1, A_2, \ldots, A_m$ are some numbers independent of Δx_1, $\Delta x_2, \ldots, \Delta x_m$ and $\alpha_1, \alpha_2, \ldots, \alpha_m$ are infinitesimal functions at $\Delta x_1 \to 0$, $x_2 \to 0$, $\ldots$, $\Delta x_m \to 0$ which are equal to zero at $\Delta x_1 =$

$= \Delta x_2 = \ldots = \Delta x_m = 0$. Relation (14.14) is called the *condition of differentiability* of a function at a given point M. Condition (14.14) of differentiability of a function can be written in another form as well. For that purpose let us consider an infinitesimal, at $\Delta x_1 \to 0, \Delta x_2 \to 0, \ldots, \Delta x_m \to 0$, function $\rho = \sqrt{\Delta x_1^2 + \Delta x_2^2 + \ldots + \Delta x_m^2}$ *, and make a note of the fact that this function vanishes only at $\Delta x_1 = \Delta x_2 = \ldots = \Delta x_m = 0$. Let us now verify that the sum $\alpha_1 \Delta x_1 + \alpha_2 \Delta x_2 + \ldots + \alpha_m \Delta x_m$ appearing on the right-hand side of relation (14.14) is an infinitesimal function of the higher order of smallness as compared to ρ. To put it otherwise, we shall ascertain that this sum is the expression $o(\rho)$. As a matter of fact, at $\rho \neq 0$ the inequality $\frac{|\Delta x_i|}{\rho} \leqslant 1$ holds true and, therefore,

$$|\alpha_1 \Delta x_1 + \alpha_2 \Delta x_2 + \ldots + \alpha_m \Delta x_m| \leqslant$$
$$\leqslant \left\{ |\alpha_1| \frac{|\Delta x_1|}{\rho} + |\alpha_2| \frac{|\Delta x_2|}{\rho} + \ldots + |\alpha_m| \frac{|\Delta x_m|}{\rho} \right\} \rho \leqslant$$
$$\leqslant \{ |\alpha_1| + |\alpha_2| + \ldots + |\alpha_m| \} \rho = o(\rho).$$

Thus we see that condition (14.14) of differentiability of a function can be written as follows:

$$\Delta u = A_1 \Delta x_1 + A_2 \Delta x_2 + \ldots + A_m \Delta x_m + o(\rho). \tag{14.15}$$

In this case the quantity $o(\rho)$ is taken to be equal to zero at $\rho = 0$.

To prove that condition (14.15) is equivalent to condition (14.14), we must verify that representation (14.15), in its turn, yields representation (14.14). For that purpose, assuming that not all the increments $\Delta x_1, \Delta x_2, \ldots, \Delta x_m$ are equal to zero**, we shall represent $o(\rho)$ in the form

$$o(\rho) = \frac{o(\rho)}{\rho} \frac{\rho^2}{\rho} = \frac{o(\rho)}{\rho} \frac{\Delta x_1^2 + \Delta x_2^2 + \ldots + \Delta x_m^2}{\rho} =$$
$$= \left[\frac{o(\rho)}{\rho} \frac{\Delta x_1}{\rho} \right] \Delta x_1 + \left[\frac{o(\rho)}{\rho} \frac{\Delta x_2}{\rho} \right] \Delta x_2 + \ldots + \left[\frac{o(\rho)}{\rho} \frac{\Delta x_m}{\rho} \right] \Delta x_m.$$

Fixing $\frac{o(\rho)}{\rho} \frac{\Delta x_i}{\rho} = \alpha_i$ and taking into account that α_i is an infinitesimal function at $\rho \to 0$ (and, hence, at $\Delta x_1 \to 0$, $\Delta x_2 \to 0$, $\ldots$, $\Delta x_m \to 0$) we shall arrive at representation (14.14).

Thus, the condition of differentiability of a function can be written both in form (14.14) and in form (14.15).

If at least one of the numbers $A_1, A_2, \ldots, A_m$ is different from zero, the sum $A_1 \Delta x_1 + A_2 \Delta x_2 + \ldots + A_m \Delta x_m$ is the *main, linear*

* In terms of geometry this function represents a distance between the points $M(x_1, x_2, \ldots, x_m)$ and $M'(x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots, x_m + \Delta x_m)$.

** If all the increments Δx_i are equal to zero, then all the terms on the right-hand side of formulas (14.14) and (14.15) are equal to zero.

with respect to the increments of the arguments, part of the increment of the differentiable function. Note that when defining the concept of differentiability of a function we did not excluded the possibility of vanishing of all the numbers $A_1, A_2, \ldots, A_m$ and, consequently, if the increment Δu of the function can be represented in form (14.14) or (14.15) at $A_1 = A_2 = \ldots = A_m = 0$, then the function is differentiable at the given point.

The following theorem holds true.

Theorem 14.9. *If the function $u = f(x_1, x_2, \ldots, x_m)$ is differentiable at point $M(x_1, x_2, \ldots, x_m)$, then there exist at this point partial derivatives with respect to all arguments, with $\frac{\partial u}{\partial x_i} = A_i$, where A_i are determined by condition (14.14) or (14.15) of differentiability of a function.*

Proof. It follows from condition (14.14) of differentiability of the function at point $M(x_1, x_2, \ldots, x_m)$ that its partial increment $\Delta_{x_i} u$ at that point is equal to $\Delta_{x_i} u = A_i \Delta x_i + \alpha_i \Delta x_i$, whence it follows that $\frac{\Delta_{x_i} u}{\Delta x_i} = A_i + \alpha_i$ and, therefore, since $\alpha_i \to 0$ at $\Delta x_i \to 0$, $\lim\limits_{\Delta x_i \to 0} \frac{\Delta_{x_i} u}{\Delta x_i} = \frac{\partial u}{\partial x_i} = A_i$.

Corollary 1. *Condition (14.15) of differentiability of a function at a given point M can be written in the following form:*

$$\Delta u = \frac{\partial u}{\partial x_1} \Delta x_1 + \frac{\partial u}{\partial x_2} \Delta x_2 + \ldots + \frac{\partial u}{\partial x_m} \Delta x_m + o(\rho)^*. \qquad (14.16)$$

Corollary 2. *If the function $u = f(x_1, x_2, \ldots, x_m)$ is differentiable at point $M(x_1, x_2, \ldots, x_m)$, then the representation of its increment Δu in form (14.14) or (14.15) is unique.* Indeed, the coefficients A_i of these representations are equal to the partial derivatives $\partial u/\partial x_i$ *at the given point M* and are therefore determined in a unique way.

Let us verify the validity of the following important property of differentiable functions.

If the function $u = f(x_1, x_2, \ldots, x_m)$ is differentiable at point $M(x_1, x_2, \ldots, x_m)$, it is continuous at that point. In fact, it follows from condition (14.14) of differentiability of a function at a point that $\lim\limits_{\substack{\Delta x_1 \to 0 \\ \Delta x_2 \to 0 \\ \cdots\cdots \\ \Delta x_m \to 0}} \Delta u = 0$ and this precisely means that the function is continuous at point M (see 14.3.1, formula (14.7)).

In the case of the function $u = f(x, y)$ of two variables the condition of differentiability can be interpreted geometrically. Let us introduce the notion of a tangent plane to the surface at point N_0. *The plane π passing through point N_0 of the surface is called a* **tangent plane** *at that point if the angle between this plane and a secant passing*

* Here all the partial derivatives $\partial u/\partial x_i$ must be taken at the given point M.

through this point N_0 and any point N_1 of the surface tends to zero as point N_1 tends to N_0 (Fig. 14.3).

If there is a tangent plane at point N_0, then, evidently, the tangent at point N_0 to any curve lying on the surface and passing through N_0 lies in the indicated plane.

Let us ascertain that the condition of differentiability of the function $u = f(x, y)$ at a given point $M_0(x_0, y_0)$ yields the existence of

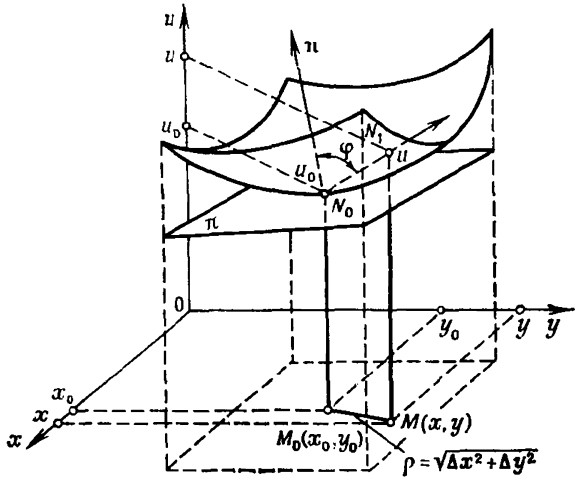


Fig. 14.3

a tangent plane to the graph S of that function at point $N_0(x_0, y_0, z_0)$. Let us put $\Delta x = x - x_0$, $\Delta y = y - y_0$, $\Delta u = u - u_0$, where $u_0 = f(x_0, y_0)$, $u = f(x, y)$. It is evident that condition (14.14) of differentiability in the case being considered can be written in the following form:

$$u - u_0 = A(x - x_0) + B(y - y_0) + \alpha\Delta x + \beta\Delta y =$$
$$= A(x - x_0) + B(y - y_0) + o(\rho),$$

where A and B are constants equal to the partial derivatives $\frac{\partial u}{\partial x}$ and $\frac{\partial u}{\partial y}$ at point M_0, and α and β are infinitesimal functions at $\Delta x \to 0$ and $\Delta y \to 0$, with $\rho = \sqrt{\Delta x^2 + \Delta y^2}$.

Let us discuss the following equation:

$$U - u_0 = A(x - x_0) + B(y - y_0).$$

It is known from analytic geometry that this equation determines, in the Cartesian system of coordinates (x, y, U), some plane π

passing through point $N_0(x_0, y_0, u_0)$ and possessing a normal vector $\mathbf{n} = \{A, B, -1\}$*.

Let us prove that the plane π is a tangent plane at point N_0 of surface S. To prove this fact, it is sufficient to ascertain that (1) the plane π passes through point N_0 of surface S, and (2) the angle φ between the normal $\mathbf{n}$ to that plane and any secant N_0N_1 tends to $\pi/2$ when point N_1 of surface S tends to point N_0. Assertion (1) is obvious. Assertion (2) requires proof. Let us calculate the cosine of the angle φ making use of the familiar formula for the cosine of the angle between two vectors. Since the coordinates of the vector $\mathbf{n}$ are equal to A, B, -1 and the coordinates of the vector $\overline{N_0N_1}$ of the secant are equal to $x - x_0$, $y - y_0$, $u - u_0$ (see Fig. 14.3), we have

$$\cos\varphi = \frac{A(x-x_0)+B(y-y_0)-(u-u_0)}{\sqrt{A^2+B^2+1}\sqrt{(x-x_0)^2+(y-y_0)^2+(u-u_0)^2}}.$$

It follows from the condition of differentiability of the function $u = (x, y)$ that

$$A(x-x_0)+B(y-y_0)-(u-u_0) = o(\rho).$$

Therefore

$$|\cos\varphi| \leqslant \frac{|o(\rho)|}{\sqrt{(x-x_0)^2+(y-y_0)^2}} = \frac{|o(\rho)|}{\rho}.$$

It follows from this formula that $\lim\limits_{\rho\to 0}\cos\varphi = 0$, i. e. $\lim\limits_{\rho\to 0}\varphi = \pi/2$. We have proved assertion (2).

Thus, interpreted geometrically, the differentiability of the function $u = f(x, y)$ at point $M_0(x_0, y_0)$ means the existence of a tangent plane to the graph of the function $u = f(x, y)$ at point $N_0(x_0, y_0, u_0)$.

Since the coefficients A and B are equal, respectively, to the partial derivatives calculated at point $M_0(x_0, y_0)$, the equation of the tangent plane can be written in the form

$$U - u_0 = \frac{\partial u}{\partial x}(x-x_0)+\frac{\partial u}{\partial y}(y-y_0). \qquad (14.17)$$

The normal vector $\mathbf{n} = \left\{\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, -1\right\}$ of the tangent plane is usually called a *normal* to the surface $u = f(x, y)$ at point $N_0(x_0, y_0, u_0)$.

Let us elucidate the sufficient conditions of differentiability of a function of several variables.

Theorem 14.10. *If the function $u = f(x_1, x_2, \ldots, x_m)$ possesses partial derivatives with respect to all arguments in some neighbourhood*

* A normal vector of a plane is any nonzero vector $\mathbf{n}$ perpendicular to that plane.

of point $M_0(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$, *with all these partial derivatives being continuous at point* M_0 *itself, the indicated function is differentiable at point* M_0.

Proof. To shorten the notation, we shall carry out the proof for the function of two variables $u = f(x, y)$. Thus suppose that both partial derivatives f_x and f_y exist in the neighbourhood of point $M_0(x_0, y_0)$ and are continuous at that point. Let us impart to the arguments x and y some increments Δx and Δy small enough for the point $M(x_0 + \Delta x, y_0 + \Delta y)$ to remain in the interior of the indicated neighbourhood of point M_0. The total increment $\Delta u = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0)$ can be written in the form

$$\Delta u = [f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0 + \Delta y)] + [f(x_0, y_0 + \Delta y) - f(x_0, y_0)].$$

The expression $[f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0 + \Delta y)]$ can be regarded as an increment of the function $f(x, y_0 + \Delta y)$ of one variable x on the closed interval $[x_0, x_0 + \Delta x]$. Since the function $u = f(x, y)$ possesses partial derivatives, the function $f(x, y_0 + \Delta y)$ is differentiable and its derivative with respect to x is a partial derivative f'_x. Applying Lagrange's formula to the indicated increment, we shall find θ_1 from the interval $0 < \theta_1 < 1$, such that

$$[f(x_0 + \Delta x, y_0 + \Delta y) - f(x, y_0 + \Delta y)] = f'_x(x_0 + \theta_1 \Delta x, y_0 + \Delta y)\,\Delta x.$$

Reasoning in the same way, we shall find that for some θ_2 from the interval $0 < \theta_2 < 1$

$$[f(x_0, y_0 + \Delta y) - f(x_0, y_0)] = f'_y(x_0, y_0 + \theta_2 \Delta y)\,\Delta y.$$

The derivatives f'_x and f'_y being continuous at point M_0, it follows that

$$f'_x(x_0 + \theta_1 \Delta x, y_0 + \Delta y) = f'_x(x_0, y_0) + \alpha,$$
$$f'_y(x_0, y_0 + \theta_2 \Delta y) = f'_y(x_0, y_0) + \beta,$$

where α and β are infinitesimal functions at $\Delta x \to 0$ and $\Delta y \to 0$. From this, with account taken of the expressions for

$$[f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0 + \Delta y] \text{ and } [f(x_0, y_0 + \Delta y) - f(x_0, y_0)]$$

mentioned above, and the expression for Δu, we find

$$\Delta u = f'_x(x_0, y_0)\,\Delta x + f'_y(x_0, y_0)\,\Delta y + \alpha \Delta x + \beta \Delta y.$$

Consequently, the function $u = f(x, y)$ is differentiable at point M_0.

In the case of a function $u = f(x_1, x_2, \ldots, x_m)$ of m variables the reasoning is the same, with the only exception that the total incre-

ment Δu of this function should be represented as the sum

$$\Delta u = f(\overset{\bullet}{x_1} + \Delta x_1, \ldots, \overset{\circ}{x_m} + \Delta x_m) - f(\overset{\circ}{x_1}, \ldots, \overset{\circ}{x_m}) =$$

$$= \sum_{k=1}^{m} [f(\overset{\bullet}{x_1}, \ldots, \overset{\circ}{x_{k-1}}, \overset{\circ}{x_k} + \Delta x_k, \overset{\circ}{x_{k+1}} + \Delta x_{k+1}, \ldots, \overset{\circ}{x_m} + \Delta x_m) -$$

$$- f(\overset{\bullet}{x_1}, \ldots, \overset{\bullet}{x_{k-1}}, \overset{\circ}{x_k}, \overset{\circ}{x_{k+1}} + \Delta x_{k+1}, \ldots, \overset{\bullet}{x_m} + \Delta x_m)].$$

The proof of the theorem is complete.

14.4.3. The concept of the differential of a function of several variables.

Definition. *The term* **differential** *du of the function* $u = f(x_1, x_2, \ldots, x_m)$ *differentiable at point* $M(x_1, x_2, \ldots, x_m)$ *is used for the main part of the increment of that function at point* M, *linear with respect to the increments of the arguments. If all the coefficients* A_i *in representation* (14.14) *of the increment of the differentiable function are equal to zero, then the differential du of the function at point* M *is assumed to be equal to zero.*

Thus, the differential du of the function $u = f(x_1, x_2, \ldots, x_m)$ differentiable at point M is the expression

$$du = A_1 \Delta x_1 + A_2 \Delta x_2 + \ldots + A_m \Delta x_m. \tag{14.18}$$

Making use of Theorem 14.9, we can, evidently, rewrite expression (14.18) for the differential du as follows:

$$du = \frac{\partial u}{\partial x_1} \Delta x_1 + \frac{\partial u}{\partial x_2} \Delta x_2 + \ldots + \frac{\partial u}{\partial x_m} \Delta x_m. \tag{14.19}$$

Let us now introduce the concept of the *differential* dx_i *of the independent variable* x_i. The differential dx_i of the independent variable x_i can be understood as any (independent of $x_1, x_2, \ldots, x_m$) number. Let us agree to assume, in what follows, this number to be equal to the increment Δx_i of the independent variable x_i. This agreement enables us to rewrite formula (14.19) in the form

$$du = \frac{\partial u}{\partial x_1} dx_1 + \frac{\partial u}{\partial x_2} dx_2 + \ldots + \frac{\partial u}{\partial x_m} dx_m. \tag{14.20}$$

It should be pointed out that formula (14.20) has been established only for the case of the arguments $x_1, x_2, \ldots, x_m$ being independent variables. However, later on, in 14.4.5, we shall prove that formula (14.20) remains valid also in the case where the arguments $x_1, x_2, \ldots, x_m$ are not independent variables but are themselves differentiable functions of some new variables.

14.4.4. Differentiation of a composite function. We shall consider here the problem of differentiation of a composite function of the

form $u = f(x_1, x_2, \ldots, x_m)$, where

$$\left.\begin{aligned} x_1 &= \varphi_1(t_1, t_2, \ldots, t_k), \\ x_2 &= \varphi_2(t_1, t_2, \ldots, t_k), \\ &\ldots\ldots\ldots\ldots \\ x_m &= \varphi_m(t_1, t_2, \ldots, t_k). \end{aligned}\right\} \tag{14.21}$$

We shall prove that under certain conditions this composite function is a differentiable function with respect to its arguments t_1, $t_2, \ldots, t_k$. In this case the partial derivatives of this composite function with respect to the arguments $t_1, t_2, \ldots, t_k$ are expressed in terms of the partial derivatives of the function $u = f(x_1, x_2, \ldots, x_m)$ and in terms of the partial derivatives of the function (14.21) by means of the following formulas:

$$\left.\begin{aligned} \frac{\partial u}{\partial t_1} &= \frac{\partial u}{\partial x_1}\frac{\partial x_1}{\partial t_1} + \frac{\partial u}{\partial x_2}\frac{\partial x_2}{\partial t_1} + \ldots + \frac{\partial u}{\partial x_m}\frac{\partial x_m}{\partial t_1}, \\ \frac{\partial u}{\partial t_2} &= \frac{\partial u}{\partial x_1}\frac{\partial x_1}{\partial t_2} + \frac{\partial u}{\partial x_2}\frac{\partial x_2}{\partial t_2} + \ldots + \frac{\partial u}{\partial x_m}\frac{\partial x_m}{\partial t_2} \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \frac{\partial u}{\partial t_k} &= \frac{\partial u}{\partial x_1}\frac{\partial x_1}{\partial t_k} + \frac{\partial u}{\partial x_2}\frac{\partial x_2}{\partial t_k} + \ldots + \frac{\partial u}{\partial x_m}\frac{\partial x_m}{\partial t_k}. \end{aligned}\right\} \tag{14.22}$$

Let us prove the following *fundamental* theorem.

Theorem 14.11. *Assume that functions* (14.21) *are differentiable at a certain point* $M(\overset{\circ}{t}_1, \overset{\circ}{t}_2, \ldots, \overset{\circ}{t}_k)$ *and the function* $u = f(x_1, x_2, \ldots, x_m)$ *is differentiable at the respective point* $N(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$, *where* $\overset{\circ}{x}_i = \varphi_i(\overset{\circ}{t}_1, \overset{\circ}{t}_2, \ldots, \overset{\circ}{t}_k)$, $i = 1, 2, \ldots, m$. *Then the composite function* $u = f(x_1, x_2, \ldots, x_m)$, *where* $x_1, x_2, \ldots, x_m$ *are specified by relations* (14.21), *is differentiable at point* M. *The partial derivatives of this composite function at point* M *are specified by formulas* (14.22) *in which all the partial derivatives* $\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \ldots, \frac{\partial u}{\partial x_m}$ *are taken at point* N *and all the partial derivatives* $\frac{\partial x_i}{\partial t_l}$ *of functions* (14.21) *with respect to the arguments* $t_1, t_2, \ldots, t_k$ *are taken at point* M.

Proof. Let us impart to the arguments $t_1, t_2, \ldots, t_k$ at point $M(\overset{\circ}{t}_1, \overset{\circ}{t}_2, \ldots, \overset{\circ}{t}_k)$ arbitrary increments $\Delta t_1, \Delta t_2, \ldots, \Delta t_k$ which are simultaneously nonzero. These increments are associated with the increments $\Delta x_1, \Delta x_2, \ldots, \Delta x_m$ of the function (14.21) at point M. The increments $\Delta x_1, \Delta x_2, \ldots, \Delta x_m$, in their turn, are associated with the increment Δu of the function $u = f(x_1, x_2, \ldots, x_m)$ at point N. Since the function $u = f(x_1, x_2, \ldots, x_m)$ is assumed to be differenti-

able at point N, the increment Δu of that function can be written as

$$\Delta u = \frac{\partial u}{\partial x_1}\Delta x_1 + \frac{\partial u}{\partial x_2}\Delta x_2 + \ldots + \frac{\partial u}{\partial x_m}\Delta x_m + \\ + \alpha_1\Delta x_1 + \alpha_2\Delta x_2 + \ldots + \alpha_m\Delta x_m, \tag{14.23}$$

where the partial derivatives $\frac{\partial u}{\partial x_1}$, $\frac{\partial u}{\partial x_2}$, ..., $\frac{\partial u}{\partial x_m}$ are taken at point N, and α_1, α_2, ..., α_m are infinitesimal functions, at $\Delta x_1 \to 0$, $\Delta x_2 \to 0$, ..., $\Delta x_m \to 0$, equal to zero at $\Delta x_1 = \Delta x_2 = \ldots = \Delta x_m = 0$. It should be emphasized that in relation (14.23) Δx_1, Δx_2, ..., Δx_m are the increments of the functions (14.21) corresponding to the chosen increments Δt_1, Δt_2, ..., Δt_k of the arguments of these functions. Since functions (14.21) are differentiable at point $M(\overset{\circ}{t}_1, \overset{\circ}{t}_2, \ldots, \overset{\circ}{t}_k)$, the increments Δx_i can be written in the following form:

$$\Delta x_i = \frac{\partial x_i}{\partial t_1}\Delta t_1 + \frac{\partial x_i}{\partial t_2}\Delta t_2 + \ldots + \frac{\partial x_i}{\partial t_k}\Delta t_k + o(\rho), \tag{14.24}$$
$$i = 1, 2, \ldots, m,$$

where the partial derivatives $\frac{\partial x_i}{\partial t_1}$, $\frac{\partial x_i}{\partial t_2}$, ..., $\frac{\partial x_i}{\partial t_k}$ are taken at point M and $\rho = \sqrt{(\Delta t_1)^2 + (\Delta t_2)^2 + \ldots + (\Delta t_k)^2}$.

We must make sure that substituting expressions (14.24) into the right-hand side of (14.23), we can reduce the increment Δu to the form 192

$$\Delta u = A_1\Delta t_1 + A_2\Delta t_2 + \ldots + A_k\Delta t_k + o(\rho), \tag{14.25}$$

where

$$A_i = \frac{\partial u}{\partial x_1}\frac{\partial x_1}{\partial t_i} + \frac{\partial u}{\partial x_2}\frac{\partial x_2}{\partial t_i} + \ldots + \frac{\partial u}{\partial x_m}\frac{\partial x_m}{\partial t_i}, \quad i = 1, 2, \ldots, k. \tag{14.26}$$

The proof of the theorem is complete since formula (14.25) establishes the fact of differentiability of a composite function and expession (14.26) is the partial derivative of the indicated composite function (see Theorem 14.9).

When we substitute expressions (14.24) into the right-hand side of (14.23), we obtain, besides the group of summands $A_1\Delta t_1 + A_2\Delta t_2 + \ldots + A_k\Delta t_k$, some other groups of summands. We must ascertain that all the other groups of summands constitute the quantity $o(\rho)$. This stems from the following considerations.

1°. *All the partial derivatives* $\partial u/\partial x_i$ *in formula* (14.23) *are taken at point* N, *that is, they are constant numbers which, being multiplied by* $o(\rho)$, *produce again the quantity* $o(\rho)$.

2°. *All* Δx_i $(i = 1, 2, \ldots, m)$ *satisfy the inequality* $|\Delta x_i| \leqslant \text{const}\ \rho$. *This follows directly from formulas* (14.24).

3°. *All* α_i *in formula* (14.23) *are infinitesimal functions at* $\rho \to 0$. In fact, all α_i are infinitesimals at $\Delta x_1 \to 0$, $\Delta x_2 \to 0$, ..., $\Delta x_m \to 0$. But all functions (14.21) are differentiable and, hence, continuous at point M. Therefore, Δx_1, Δx_2, ..., Δx_m tends to zero as $\rho \to 0$.

4°. *Each product* $\alpha_i \Delta x_i$ *constitutes the quantity* $o(\rho)$. This follows directly from 2° and 3°. We have proved the theorem.

Remark. Let us consider a significant particular case where functions (14.21) depend on one argument t. Then we have a composite function of one variable t: $u = f(x_1, x_2, \ldots, x_m)$, where $x_i = \varphi_i(t)$. The derivative du/dt of this composite function is determined by the following formula:

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial u}{\partial x_2}\frac{\partial x_2}{\partial t} + \ldots + \frac{\partial u}{\partial x_m}\frac{\partial x_m}{\partial t}. \tag{14.27}$$

Let us apply formula (14.27) to prove *Euler's theorem on homogeneous functions.*

The function $u = f(x_1, x_2, \ldots, x_m)$ defined on the set $\{M\}$ is called a homogeneous function of degree p on this set if for every point $M(x_1, x_2, \ldots, x_m)$ of the set $\{M\}$ and for every number t for which point $N(tx_1, tx_2, \ldots, tx_m)$ belongs to the set $\{M\}$, *there holds an equality*

$$f(tx_1, tx_2, \ldots, tx_m) = t^p f(x_1, x_2, \ldots, x_m). \tag{14.28}$$

Theorem 14.12 (Euler's theorem on homogeneous functions). *If* $u = f(x_1, x_2, \ldots, x_m)$ *is a homogeneous differentiable function of degree* p *in some domain* $\{M\}$, *then the equality*

$$\frac{\partial u}{\partial x_1}x_1 + \frac{\partial u}{\partial x_2}x_2 + \ldots + \frac{\partial u}{\partial x_m}x_m = pu \tag{14.29}$$

holds at every point $M(x_1, x_2, \ldots, x_m)$ *of the domain* $\{M\}$.

Proof. Suppose $M_0(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$ is an arbitrary point from the domain $\{M\}$. Let us discuss the composite function $u = f(x_1, x_2, \ldots, x_m)$ where $x_i = t\overset{\circ}{x}_i$ $(i = 1, 2, \ldots, m)$, i.e. the function $u = f(t\overset{\circ}{x}_1, t\overset{\circ}{x}_2, \ldots, t\overset{\circ}{x}_m)$. Since at $t = 1$ the functions $x_i = t\overset{\circ}{x}_i$ are differentiable and the function $u = f(x_1, x_2, \ldots, x_m)$ is differentiable at the corresponding point M_0. Theorem 14.11 and the remark to that theorem enable us to calculate the derivative $\frac{du}{dt}$ of the indicated composite function at point $t = 1$ by formula (14.27). Since $\frac{dx_i}{dt} = \overset{\circ}{x}_i$, it follows that

$$\left.\frac{du}{dt}\right|_{t=1} = \frac{\partial u}{\partial x_1}\overset{\circ}{x}_1 + \frac{\partial u}{\partial x_2}\overset{\circ}{x}_2 + \ldots + \frac{\partial u}{\partial x_m}\overset{\circ}{x}_m, \tag{14.30}$$

where the derivatives $\frac{\partial u}{\partial x_i}$ are taken at point M_0. On the other hand, by virtue of (14.28), the composite function in question can be re-

presented as follows:

$$u = f(t\overset{\circ}{x}_1, t\overset{\circ}{x}_2, \ldots, t\overset{\circ}{x}_m) = t^p f(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m). \tag{14.31}$$

It follows from (14.31) that $\frac{du}{dt} = pt^{p-1} f(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$, i.e.

$$\left.\frac{du}{dt}\right|_{t=1} = pf(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m) = pu. \tag{14.32}$$

Comparing (14.30) and (14.32), we obtain relation (14.29) for point M_0. Since point M_0 is an arbitrary point of the domain $\{M\}$, the proof of the theorem is complete.

14.4.5. Invariance of the form of the first differential. In 14.4.3 we have introduced the concept of the first differential du of a function of several variables and established the fact that when the arguments $x_1, x_2, \ldots, x_m$ are independent variables, the differential du can be represented in the form

$$du = \frac{\partial u}{\partial x_1} dx_1 + \frac{\partial u}{\partial x_2} dx_2 + \ldots + \frac{\partial u}{\partial x_m} dx_m. \tag{14.20}$$

Here we shall prove that formula (14.20) is universal and is valid even in the case where the arguments $x_1, x_2, \ldots, x_m$ are themselves differentiable functions of the new variables $t_1, t_2, \ldots, t_k$. This property of the first differential is usually called the *invariance of its form.*

Thus, let us assume that the arguments $x_1, x_2, \ldots, x_m$ of the function $u = f(x_1, x_2, \ldots, x_m)$ are the functions $x_i = \varphi_i(t_1, t_2, \ldots, t_k)$ differentiable at point $A(\overset{\circ}{t}_1, \overset{\circ}{t}_2, \ldots, \overset{\circ}{t}_k)$ and the function $u = f(x_1, x_2, \ldots, x_m)$ itself is differentiable at point $B(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$, where $\overset{\circ}{x}_i = \varphi_i(\overset{\circ}{t}_1, \overset{\circ}{t}_2, \ldots, \overset{\circ}{t}_k)$. In such a case we can consider u as a composite function with respect to the arguments $t_1, t_2, \ldots, t_k$ which, in accordance with Theorem 14.11, is differentiable at point A. Therefore, the differential du of this composite function can be represented in the form

$$du = \frac{\partial u}{\partial t_1} dt_1 + \frac{\partial u}{\partial t_2} dt_2 + \ldots + \frac{\partial u}{\partial t_k} dt_k, \tag{14.33}$$

where $\frac{\partial u}{\partial t_i}$ can be determined from relations (14.22). Substituting $\frac{\partial u}{\partial t_i}$ from (14.22) into (14.33) and combining the coefficients in $\frac{\partial u}{\partial x_i}$, we obtain

$$du = \frac{\partial u}{\partial x_1}\left(\frac{\partial x_1}{\partial t_1} dt_1 + \frac{\partial x_1}{\partial t_2} dt_2 + \ldots + \frac{\partial x_1}{\partial t_k} dt_k\right) + \ldots$$
$$\ldots + \frac{\partial u}{\partial x_m}\left(\frac{\partial x_m}{\partial t_1} dt_1 + \frac{\partial x_m}{\partial t_2} dt_2 + \ldots + \frac{\partial x_m}{\partial t_k} dt_k\right).$$

It remains to take a note of the fact that in the last relation the coefficient in $\partial u/\partial x_i$ is equal to the differential dx_i of the function $x_i = \varphi_i(t_1, t_2, \ldots, t_k)$. We shall obtain formula (14.20) for the differential du of the composite function, the differentials dx_i in this formula being the differentials of the functions $x_i = \varphi_i(t_1, t_2, \ldots, t_k)$. We have thus established the invariance of the form of the first differential.

The property of the invariance of the form of the first differential makes it possible to establish the following *differentiation rules.* Suppose u and v are differentiable functions of some variables. Then

$$d(cu) = c\,du \ (c = \text{const}),$$

$$d(u \pm v) = du \pm dv, \quad d(uv) = u\,dv + v\,du,$$

$$d\left(\frac{u}{v}\right) = \frac{v\,du - u\,dv}{v^2}.$$

(In the last of the formulas written above v does not vanish.)

Let us prove, for example, the validity of the third of the indicated formulas. We shall consider the function $w = uv$ of two variables u and v. The differential dw of this function is

$$dw = \frac{\partial w}{\partial u}\,du + \frac{\partial w}{\partial v}\,dv.$$

Since $\partial w/\partial u = v$ and $\partial w/\partial v = u$, it follows that $dw = u\,dv + v\,du$. By virtue of the invariance of the form of the first differential, the expression $u\,dv + v\,du$ is the differential of the function uv also in the case where u and v are themselves differentiable functions of some variables.

14.4.6. Directional derivative. Gradient. Let us assume that the function $u = f(x, y, z)$ of three variables x, y and z is given in some neighbourhood of point $M_0(x_0, y_0, z_0)$. Let us consider a direction determined by a unit vector **a** with the coordinates $\cos\alpha$, $\cos\beta$, $\cos\gamma$*. Let us draw through point M_0 an axis **l** whose direction coincides with the direction of vector **a**, take an arbitrary point $M(x, y, z)$ on that axis and designate by l the magnitude of the directed segment M_0M of that axis**. It is known from analytic geometry that the coordinates x, y, z of point M are specified by the equalities

$$x = x_0 + l\cos\alpha, \quad y = y_0 + l\cos\beta, \quad z = z_0 + l\cos\gamma. \quad (14.34)$$

On the indicated axis **l** the function $u = f(x, y, z)$ is, evidently, a composite function of one variable quantity l. *If this function posses-*

* It is known from analytic geometry that if the unit vector **a** forms angles α, β, γ with the axes of coordinates, the coordinates of **a** are $\cos\alpha$, $\cos\beta$, $\cos\gamma$.

** The magnitude l of the directed segment M_0M of axis **l** is the number equal to its length taken with the plus sign if the direction of the segment coincides with that of axis **l**, and with the minus sign if the direction of that segment is opposite to the direction of axis **l**.

ses at point $l = 0$ a derivative with respect to the variable l, then this derivative is called a **derivative in the direction l** *of the function $u = f(x, y, z)$ at point M_0 and is designated as $\partial u/\partial t$.* In accordance with the remark to Theorem 14.11, in the case of differentiability of the function $u = f(x, y, z)$ at point M_0 the derivative $\partial u/\partial t$ can be calculated by formula (14.27) in which the argument t must be replaced by l. Thus we have

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x}\frac{\partial x}{\partial l} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial l} + \frac{\partial u}{\partial z}\frac{\partial z}{\partial l}.$$

Since $\frac{dx}{dl} = \cos\alpha$, $\frac{dy}{dl} = \cos\beta$, $\frac{dz}{dl} = \cos\gamma$, it follows from the last formula that

$$\frac{\partial u}{\partial l} = \frac{\partial u}{\partial x}\cos\alpha + \frac{\partial u}{\partial y}\cos\beta + \frac{\partial u}{\partial z}\cos\gamma. \tag{14.35}$$

Let us now introduce the concept of the *gradient* of the function $u = f(x, y, z)$ differentiable at point $M_0(x_0, y_0, z_0)$.

The **gradient** *of the function $u = f(x, y, z)$ at point M_0 is the vector designated by the symbol grad u and having the coordinates equal, respectively, to the derivatives $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$, $\frac{\partial u}{\partial z}$ taken at point M_0.* Thus we have

$$\operatorname{grad} u = \left\{\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial u}{\partial z}\right\}. \tag{14.36}$$

Making use of the notion of the gradient of a function and taking into account that the vector **a**, which determines the direction of the axis **l**, has the coordinates $\cos\alpha$, $\cos\beta$, $\cos\gamma$, we shall represent expression (14.35) for the derivative $\partial u/\partial l$ with respect to the direction **l** as a scalar product of the vectors grad u and **a**:

$$\frac{\partial u}{\partial l} = \mathbf{a}\operatorname{grad} u^*. \tag{14.37}$$

Let us show that *the gradient of the function $u = f(x, y, z)$ at point M_0 characterizes the direction and the magnitude of the maximum growth of the function at point M_0*, namely, let us verify that the derivative of the function u at point M_0 along the direction determined by the gradient of that function at the indicated point has the maximum value as compared to the derivative along any other direction at point M_0, the value of the indicated derivative being equal to $|\operatorname{grad} u|$, that is, to the length of the vector grad u. Let us rewrite

---

* Recall that a scalar product of two vectors defined as the product of their absolute values (lengths) by the cosine of the angle between them in the case where the vectors are specified by coordinates is equal to the sum of the products of the like coordinates of the vectors.

formula (14.37) in the form

$$\frac{\partial u}{\partial l} = |\mathbf{a}|\,|\operatorname{grad} u|\cos\varphi,$$

where φ is the angle between the vectors **a** and grad u. Since

$|\mathbf{a}| = 1$, we have $\frac{\partial u}{\partial l} = |\operatorname{grad} u|\cos\varphi$.

It follows from the last formula that the maximum value $\left(\frac{\partial u}{\partial l}\right)_{\max}$ of the directional derivative will be at $\cos\varphi = 1$, that is in the case where the direction of the vector **a** coincides with that of grad **u**, with $\left(\frac{\partial u}{\partial l}\right)_{\max} = |\operatorname{grad} u|$.

To elucidate the geometrical meaning of the vector grad u, let us introduce the concept of the *level surface* of the function $u = f(x, y, z)$.

We shall use the term the level surface of the function $u = f(x, y, z)$ for every surface at which the function $u = f(x, y, z)$ retains a constant value, $f(x, y, z) = c = \text{const}$.

It is easy to verify that the vector grad u at the given point $M_0(x_0, y_0, z_0)$ is orthogonal to the level surface of the function $u = f(x, y, z)$ which passes through the given point M_0.

Remark. In the case of the function $u = f(x, y)$ of two variables, x and y, the unit vector **a**, determining the direction at point M_0, has the coordinates $\cos\alpha$ and $\sin\alpha$. Therefore, in this case formula (14.35) assumes the form

$$\frac{\partial u}{\partial l} = \frac{\partial u}{\partial x}\cos\alpha + \frac{\partial u}{\partial y}\sin\alpha.$$

It should be pointed out that in the case of a function of two variables the gradient of the differentiable function $u(x, y)$ is defined as a vector having the coordinates $\frac{\partial u}{\partial x}$ and $\frac{\partial u}{\partial y}$. Formula (14.37) is, obviously, valid in the case of two variables as well. For the function $u = f(x_1, x_2, \ldots, x_m)$ of m variables $x_1, x_2, \ldots, x_m$, the directional derivative and the gradient are determined in the same way. In fact, the derivative $\frac{\partial u}{\partial l}$ at point $M_0(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$ in the direction **l**, which is specified by the unit vector $\mathbf{a} = \{\cos\alpha_1, \cos\alpha_2, \ldots, \cos\alpha_m\}$*, is defined as a derivative with respect to l of the composite function $u = f(x_1, x_2, \ldots, x_m)$, where $x_1 = \overset{\circ}{x}_1 + l\cos\alpha_1$, $x_2 = \overset{\circ}{x}_2 + l\cos\alpha_2$, $\ldots$, $x_m = \overset{\circ}{x}_m + l\cos\alpha_m$. In the case where $u = f(x_1, x_2, \ldots, x_m)$ is a differentiable function, the

* In the analytic geometry of the m-dimensional Euclidean space the unit vector a is defined as the vector with the coordinates $\cos\alpha_1, \cos\alpha_2, \ldots, \cos\alpha_m$, where $\cos^2\alpha_1 + \cos^2\alpha_2 + \ldots + \cos^2\alpha_m = 1$.

following formula holds true for the directional derivative:

$$\frac{\partial u}{\partial l} = \frac{\partial u}{\partial x_1}\cos\alpha_1 + \frac{\partial u}{\partial x_2}\cos\alpha_2 + \ldots + \frac{\partial u}{\partial x_m}\cos\alpha_m.$$

The gradient of a function at the given point $M_0(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$ is the vector symbolized as grad u and having the coordinates $\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \ldots, \frac{\partial u}{\partial x_m}$, the indicated derivatives being taken at point M_0. Formula (14.37) is valid for the directional derivative of the differentiable function $u = f(x_1, x_2, \ldots, x_m)$.

14.5. PARTIAL DERIVATIVES AND DIFFERENTIALS OF HIGHER ORDERS

14.5.1. Partial derivatives of higher orders. Assume that the partial derivative $\partial u/\partial x_i$ with respect to the argument x_i of the function $u = f(x_1, x_2, \ldots, x_m)$, defined in the domain $\{M\}$ exists at every point of the domain $\{M\}$. In that case the indicated partial derivative is a function of the variables $x_1, x_2, \ldots, x_m$ also defined in the domain $\{M\}$. This function $\partial u/\partial x_i$ may happen to possess a partial derivative with respect to the argument x_k at some point M of the domain $\{M\}$. Then this partial derivative with respect to the argument x_k is called the second partial derivative or the second-order derivative of the function $u = f(x_1, x_2, \ldots, x_m)$ at point M first with respect to the argument x_i and then with respect to the argument x_k and is designated by one of the symbols

$$\frac{\partial^2 u}{\partial x_k\,\partial x_i}, \quad f''_{x_i x_k}, \quad u''_{x_i x_k}.$$

In this case, if $i \neq k$, the partial derivative $\partial^2 u/(\partial x_k \partial x_i)$ is called a *mixed* partial derivative of the second order. The concept of the second partial derivative having been introduced, we can successively introduce the notions of the third partial derivative, the fourth partial derivative, etc. If we assume that we have introduced the notion of the $(n-1)$th partial derivative of the function $u = f(x_1, x_2, \ldots, x_m)$ with respect to the arguments $x_{i_1}, x_{i_2}, \ldots, x_{i_{n-1}}$ (whose separate or even all numbers may coincide) and that the $(n-1)$th partial derivative possesses at point M a partial derivative with respect to the argument x_{i_n}, the indicated partial derivative is called the nth partial derivative (or a partial derivative of the nth order) of the function $u = f(x_1, x_2, \ldots, x_m)$ at point M with respect to the arguments $x_{i_1}, x_{i_2}, \ldots, x_{i_{n-1}}, x_{i_n}$. Thus we use the *method of induction* to introduce the notion of the nth partial derivative, passing from the first partial derivative to the next ones. The relation specifying the nth partial derivative with respect to the arguments

$x_{i_1}, x_{i_2}, \ldots, x_{i_{n-1}}, x_{i_n}$ has the form

$$\frac{\partial^n u}{\partial x_{i_n} \partial x_{i_{n-1}} \ldots \partial x_{i_2} \partial x_{i_1}} = \frac{\partial}{\partial x_{i_n}} \left(\frac{\partial^{n-1} u}{\partial x_{i_{n-1}} \ldots \partial x_{i_2} \partial x_{i_1}} \right).$$

Now if the indices $i_1, i_2, \ldots, i_n$ do not all coincide, the partial derivative $\frac{\partial^n u}{\partial x_{i_n} \ldots \partial x_{i_2} \partial x_{i_1}}$ is called a *mixed* partial derivative of the nth order. Since the partial derivative of the function with respect to the argument x_i is defined as an ordinary derivative of a function of one variable x_i with the values of the rest of the variables being fixed, the method of calculation of the partial derivatives of higher orders presupposes only the ability to calculate ordinary derivatives of the first order. By way of example, we shall calculate the partial derivatives of the second order of the function $u = \arctan \frac{x}{y}$. We have

$$\frac{\partial u}{\partial x} = \frac{y}{x^2+y^2}, \qquad \frac{\partial u}{\partial y} = -\frac{x}{x^2+y^2},$$

$$\frac{\partial^2 u}{\partial x^2} = -\frac{2xy}{(x^2+y^2)^2}, \quad \frac{\partial^2 u}{\partial x \partial y} = \frac{x^2-y^2}{(x^2+y^2)^2},$$

$$\frac{\partial^2 u}{\partial y \partial x} = \frac{x^2-y^2}{(x^2+y^2)^2}, \qquad \frac{\partial^2 y}{\partial y^2} = \frac{2xy}{(x^2+y^2)}.$$

In the example we have considered the mixed partial derivatives $\frac{\partial^2 u}{\partial y \partial x}$ and $\frac{\partial^2 u}{\partial x \partial y}$ are equal to each other. In a general case, the values of **mixed** derivatives depend on the order the successive differentiations are performed. Let us verify, for instance, that the mixed partial derivatives $\frac{\partial^2 u}{\partial y \partial x}$ and $\frac{\partial^2 u}{\partial x \partial y}$ of the function

$$u = \begin{cases} xy \dfrac{x^2-y^2}{x^2+y^2} & \text{for } x^2+y^2 \neq 0, \\ 0 & \text{for } x^2+y^2 = 0 \end{cases}$$

at point (0, 0) exist but are not equal to each other. Indeed,

$$\frac{\partial u}{\partial x} = \begin{cases} \dfrac{y(x^4-y^4+4x^2y^2)}{(x^2+y^2)^2} & \text{for } x^2+y^2 \neq 0. \\ 0 & \text{for } x^2+y^2 = 0. \end{cases}$$

Therefore,

$$\left.\frac{\partial^2 u}{\partial y \partial x}\right|_{x=0,\, y=0} = \lim_{y \to 0} \frac{\left.\frac{\partial u}{\partial x}\right|_{x=0,\, y \neq 0} - \left.\frac{\partial u}{\partial x}\right|_{x=0,\, y=0}}{y} = -1.$$

Carrying out similar calculations, we obtain $\left.\frac{\partial^2 u}{\partial x \partial y}\right|_{x=0,\, y=0} = 1$. Thus we have $\frac{\partial^2 u}{\partial x \partial y} \neq \frac{\partial^2 u}{\partial y \partial x}$ at point (0, 0). Let us elucidate the

sufficient conditions of independence of the values of mixed partial derivatives of the order the successive differentiations are performed. Let us first introduce the notion of the n-times differentiable function of several variables. *The function* $u=f(x_1, x_2, \ldots, x_m)$ *is said to be* n **times differentiable** *at point* $M_0(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$ *if all the partial derivatives of the* $(n-1)$*th order of that function are differentiable functions at point* M_0. Take a note of the following assertion. *For the function* $u=f(x_1, x_2, \ldots, x_m)$ *to be* n *times differentiable at point* $M_0(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$ *it is sufficient that all its derivatives of the* n*th order should be continuous at point* M_0. The validity of this assertion follows from the definition of differentiability of a function and Theorem 14.10 on the sufficient conditions of differentiability.

Theorem 14.13. *Assume the function* $u=f(x, y)$ *to be twice differentiable at point* $M_0(x_0, y_0)$. *Then the partial derivatives* f''_{xy} *and* f''_{yx} *are equal at that point.*

Proof. Since the function $u=f(x, y)$ is twice differentiable at point $M_0(x_0, y_0)$, the partial derivatives f'_x and f'_y are defined in some neighbourhood of point M_0 and are differentiable functions at that point. Let us consider the expression

$$\Phi = f(x_0+h, y_0+h) - f(x_0+h, y_0) - f(x_0, y_0+h) + f(x_0, y_0), \tag{14.38}$$

where h is any infinitesimal such that point $M(x_0+h, y_0+h)$ is in the interior of the indicated neighbourhood of point M_0. The expression Φ can be regarded as an increment $\Delta\varphi = \varphi(x_0+h) - \varphi(x_0)$ of the function $\varphi(x) = f(x, y_0, +h) - f(x, y_0)$, with respect to one variable x, differentiable on the closed interval $[x_0, x_0+h]$. Therefore, in accordance with Lagrange's formula, denoting by θ some number from the interval $0<\theta<1$, we can write

$$\Phi = \Delta\varphi = \varphi'(x_0+\theta h)\,h = [f'_x(x_0+\theta h, y_0+h) - f'_x(x_0+\theta h, y_0)]\,h = \{[f'_x(x_0+\theta h, y_0+h) - f'_x(x_0, y_0)] - [f'_x(x_0+\theta h, y_0) - f'_x(x_0, y_0)]\}\,h. \tag{14.39}$$

The partial derivative f'_x being a differentiable function at point M_0, it follows that

$$[f'_x(x_0+\theta h, y_0+h) - f'_x(x_0, y_0)] = f''_{xx}(x_0, y_0)\,\theta h + f''_{xy}(x_0, y_0)\,h + \alpha_1\theta h + \beta_1 h,$$

$$[f'_x(x_0+\theta h, y_0) - f'_x(x_0, y_0)] = f''_{xx}(x_0, y_0)\,\theta h + \alpha_2\theta h,$$

where α_1, β_1, and α_2 are infinitesimal functions at $h \to 0$. Substituting the expressions obtained for $[f'_x(x_0 + \theta h, y_0 + h) - f'_x(x_0, y_0)]$ and $[f'_x(x_0 + \theta h, y_0) - f'_x(x_0, y_0)]$ into formula (14.39), we obtain

$$\Phi = [f''_{xy}(x_0, y_0) + \alpha] h^2, \tag{14.40}$$

where $\alpha = \alpha_1\theta + \beta_1 - \alpha_2\theta$ is an infinitesimal function at $h \to 0$. On the other hand, the expression for Φ specified by relation (14.38) can be regarded as an increment $\Delta\psi = \psi(y_0 + h) - \psi(y_0)$ of the function $\psi(y) = f(x_0 + h, y) - f(x_0, y)$ differentiable on the closed interval $[y_0, y_0 + h]$. Applying Lagrange's formula and taking into account the differentiability of the partial derivative f'_y at point M_0, we shall obtain, using the same reasoning as before, the following expression for Φ:

$$\Phi = [f''_{yx}(x_0, y_0) + \beta] h^2, \tag{14.41}$$

where β is an infinitesimal function at $h \to 0$. Equating the right-hand sides of relations (14.40) and (14.41) to each other and cancelling h^2 from both sides of the resulting equality, we find that $f''_{xy}(x_0, y_0) + \alpha = f''_{yx}(x_0, y_0) + \beta$. Since α and β are infinitesimal functions at $h \to 0$, the last equality yields $f''_{xy}(x_0, y_0) = f''_{yx}(x_0, y_0)$. We have proved the theorem.

Remark. Theorem 14.13 states that the equality $f''_{xy} = f''_{yx}$ holds at the given point $M_0(x_0, y_0)$ if f'_x and f'_y are differentiable at that point. The differentiability of f'_x and f'_y at point M_0 implies the existence *at that point of all* the second-order partial derivatives. However, the equality $f''_{xy} = f''_{yx}$ holds true even under the condition of the existence only of the derivatives f''_{xy} and f''_{yx} but with the additional requirement of continuity of those derivatives at the point in question. Namely, the following assertion holds true.

Assume that the function $u = f(x, y)$ possesses partial derivatives f'_x, f'_y, f''_{xy}, f''_{yx} in some neighbourhood of point $M_0(x_0, y_0)$. Suppose, in addition, that the derivatives f''_{xy} and f''_{yx} are continuous at point M_0. Then at that point $f''_{xy} = f''_{yx}$.

To prove the assertion, let us make use of the expression for Φ specified by relations (14.38). It follows from (14.39) that Φ is the difference in the values of the function $f'_x(x, y)$ at points $(x_0 + \theta h, y_0 + h)$ and $(x_0 + \theta h, y_0)$ multiplied by h. Applying to this difference Lagrange's formula of finite increments with respect to the variable y on the closed interval $[y_0, y_0 + h]$, we obtain

$$\Phi = f''_{xy}(x_0 + \theta h, y_0 + \theta_1 h) h^2, \text{ where } 0 < \theta_1 < 1.$$

The derivative f''_{xy} being continuous at point $M_0(x_0, y_0)$, the last equality yields

$$\Phi = [f''_{xy}(x_0, y_0) + \alpha(h)] h^2$$

where $\alpha(h) \to 0$ as $h \to 0$.

On the other hand, this same quantity Φ is the difference in the values of the function $f'_y(x, y)$ at points $(x_0 + h, y_0 + \theta_2 h)$ and $(x_0, y_0 + \theta_2 h)$ multiplied

by h. Applying to this difference Lagrange's formula of finite increments with respect to the variable x on the closed interval $[x_0, x_0 + h]$ and taking into account the continuity of f''_{yx} at point $M_0(x_0, y_0)$, we obtain

$$\Phi = [f''_{yx}(x_0, y_0) + \beta(h)]\, h^2,$$

where $\beta(h) \to 0$ as $h \to 0$.

Equating the last two expressions for Φ to each other and reasoning as we did at the end of the proof of Theorem 14.13, we shall ascertain the validity of the required equality

$$f''_{xy}(x_0, y_0) = f''_{yx}(x_0, y_0).$$

Let us now prove the theorem on the independence of the value of any nth order mixed partial derivative of the order in which successive differentiations are carried out.

Theorem 14.14. *Suppose the function $u = f(x_1, x_2, \ldots, x_m)$ is n times differentiable at point $M_0(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$. Then at that point the value of any nth order mixed partial derivative does not depend on the order of successive differentiations.*

Proof. It is, evidently, sufficient to prove the independence of the value of any nth mixed derivative from the order in which *two consecutive* differentiations are carried out. To put it otherwise, it is sufficient to prove the equality

$$\frac{\partial^n u}{\partial x_{i_n} \ldots \partial x_{i_{k+1}} \partial x_{i_k} \ldots \partial x_{i_1}} = \frac{\partial^n u}{\partial x_{i_n} \ldots \partial x_{i_k} \partial x_{i_{k+1}} \ldots \partial x_{i_1}}. \tag{14.42}$$

Let us consider the function $\dfrac{\partial^{k-1} u}{\partial x_{i_{k-1}} \ldots \partial x_{i_1}}$. This is a twice differentiable function with respect to the variables x_{i_k} and $x_{i_{k+1}}$. Therefore, by virtue of Theorem 14.13,

$$\frac{\partial^{k+1} u}{\partial x_{i_{k+1}} \partial x_{i_k} \partial x_{i_{k-1}} \ldots \partial x_{i_1}} = \frac{\partial^{k+1} u}{\partial x_{i_k} \partial x_{i_{k+1}} \partial x_{i_{k-1}} \ldots \partial x_{i_1}}.$$

And this implies the validity of equality (14.42). The proof of the theorem is complete.

Note that in the case of the function $u = f(x_1, x_2, \ldots, x_m)$ being n times differentiable at point M_0, its any partial derivative of the nth order can be written in the form

$$\frac{\partial^n n}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \ldots \partial x_m^{\alpha_m}}$$

where $\alpha_1, \alpha_2, \ldots, \alpha_m$ are integers satisfying the conditions $0 \leqslant \alpha_i \leqslant n$, $\alpha_1 + \alpha_2 + \ldots + \alpha_m = n$.

14.5.2. Differentials of higher orders. We have earlier employed the symbols $dx_1, dx_2, \ldots, dx_m$ and du to designate the differentials of the arguments of the function $u = f(x_1, x_2, \ldots, x_m)$ and the differential of the function itself, respectively.

Now we have to use other symbols for the same purposes. In particular, we shall denote the differentials of the arguments of the function $u = f(x_1, x_2, \ldots, x_m)$ and the differential of the function itself by the symbols $\delta x_1, \delta x_2, \ldots, \delta x_m$ and δu, respectively. In these designations the expression for the first differential of the

function (14.20), invariant in its form (see 4.5), will have the form

$$\delta u = \frac{\partial u}{\partial x_1}\delta x_1 + \frac{\partial u}{\partial x_2}\delta x_2 + \ldots + \frac{\partial u}{\partial x_m}\delta x_m.$$

Returning to the notation used before, we shall consider expression (14.20) for the first differential of the function $u = f(x_1, x_2, \ldots, x_m)$ differentiable at the given point $M(x_1, x_2, \ldots, x_m)$:

$$du = \frac{\partial u}{\partial x_1}dx_1 + \frac{\partial u}{\partial x_2}dx_2 + \ldots + \frac{\partial u}{\partial x_m}dx_m. \qquad (14.20)$$

Let us assume that the quantity appearing on the right-hand side of (14.20) is a function of the arguments $x_1, x_2, \ldots, x_m$ differentiable at the given point $M(x_1, x_2, \ldots, x_m)$. To do that, it is sufficient to require that the function $u = f(x_1, x_2, \ldots, x_m)$ be twice differentiable at the given point $M(x_1, x_2, \ldots, x_m)$ and the arguments $x_1, x_2, \ldots, x_m$ be either independent variables or twice differentiable functions of some independent variables.

Under these assumptions, we can consider the differential

$$\delta(du) = \delta\left[\sum_{k=1}^{m}\frac{\partial u}{\partial x_k}dx_k\right]$$

of the quantity (14.20).

Definition 1. *The value $\delta(du)$ of the differential of the first differential* (14.20) *taken at $\delta x_1 = dx_1$, $\delta x_2 = dx_2$, ..., $\delta x_m = dx_m$ is said to be the* **second differential** *of the function $u = f(x_1, x_2, \ldots, x_m)$ (at the given point $M(x_1, x_2, \ldots, x_m)$) and is symbolized as d^2u.*

Thus, in accordance with the definition*,

$$d^2u = \delta(du)\Big|_{\substack{\delta x_1 = dx_1\\ \delta x_2 = dx_2\\ \ldots\\ \delta x_m = dx_m}} = \left\{\delta\left[\sum_{k=1}^{m}\frac{\partial u}{\partial x_k}dx_k\right]\right\}\Bigg|_{\substack{\delta x_1 = dx_1\\ \delta x_2 = dx_2\\ \ldots\\ \delta x_m = dx_m}}$$

We shall use the **method of induction** to introduce the differential d^nu of any order n.

Let us assume that we have already introduced the differential $d^{n-1}u$ of the order $n-1$ and that the function $u = f(x_1, x_2, \ldots, x_m)$ is n times differentiable at the given point $M(x_1, x_2, \ldots, x_m)$ and its arguments $x_1, x_2, \ldots, x_m$ are either independent variables or n times differentiable functions of some independent variables $t_1, t_2, \ldots, t_k$.

* The symbol $\left\{\ \right\}\Big|_{\substack{\delta x_1 = dx_1\\ \delta x_2 = dx_2,\\ \ldots\\ \delta x_m = dx_m}}$ signifies that in the expression in braces we have to put $\delta x_1 = dx_1$, $\delta x_2 = dx_2, \ldots, \delta x_m = dx_m$.

Definition 2. *The value* $\delta(d^{n-1}u)$ *of the differential of the* $(n-1)$*th differential* $d^{n-1}u$ *taken at* $\delta x_1 = dx_1$, $\delta x_2 = dx_2$, . . ., $\delta x_m = dx_m$ *is said to be the* ***n*th differential** *of the function* $u = f(x_1, x_2, \ldots, x_m)$ *(at the given point* $M(x_1, x_2, \ldots, x_m)$*) and is symbolized as* $d^n u$.

Thus, in accordance with the definition

$$d^n u = \delta(d^{n-1}u)\Big|_{\substack{\delta x_1 = dx_1 \\ \delta x_2 = dx_2 \\ \cdots \\ \delta x_m = dx_m}}.$$

Two cases should be distinguished in calculating the second and successive integrals: (1) the arguments $x_1, x_2, \ldots, x_m$ are independent variables, (2) the arguments $x_1, x_2, \ldots, x_m$ are functions of some independent variables $t_1, t_2, \ldots, t_k$ differentiable the corresponding number of times.

We shall begin with the first case. If $x_1, x_2, \ldots, x_m$ are **independent** variables, we have the right to assume that $dx_1, dx_2, \ldots, dx_m$ do not depend on $x_1, x_2, \ldots, x_m$.

Each differential dx_k can be taken to be equal to one and the same increment Δx_k for all the points $M(x_1, x_2, \ldots, x_m)$. Then we shall obtain

$$\delta(dx_k) = \sum_{i=1}^{m} \frac{\partial(dx_k)}{\partial x_i}\delta x_i = 0.$$

The last relation and the rules of differentiation established at the end of 14.4.5 enable us to write the following chain of equalities for the function $u = f(x_1, x_2, \ldots, x_m)$ twice differentiable at the given point M:

$$d^2u = \delta(du)\Big|_{\substack{\delta x_1 = dx_1 \\ \cdots \\ \delta x_m = dx_m}} = \delta\left[\sum_{k=1}^{m}\frac{\partial u}{\partial x_k}dx_k\right]\Bigg|_{\substack{\delta x_1 = dx_1 \\ \cdots \\ \delta x_m = dx_m}} =$$

$$= \sum_{k=1}^{m}\delta\left[\frac{\partial u}{\partial x_k}dx_k\right]\Bigg|_{\substack{\delta x_1 = dx_1 \\ \cdots \\ \delta x_m = dx_m}} = \sum_{k=1}^{m}\left\{dx_k\cdot\delta\left(\frac{\partial u}{\partial x_k}\right) + \right.$$

$$\left. + \frac{\partial u}{\partial x_k}\delta(dx_k)\right\}\Bigg|_{\substack{\delta x_1 = dx_1 \\ \cdots \\ \delta x_m = dx_m}} = \sum_{k=1}^{m}\left\{dx_k\sum_{i=1}^{m}\frac{\partial}{\partial x_i}\left(\frac{\partial u}{\partial x_k}\right)\delta x_i\right\}\Bigg|_{\substack{\delta x_1 = dx_1 \\ \cdots \\ \delta x_m = dx_m}} =$$

$$= \left[\sum_{k=1}^{m}\sum_{i=1}^{m}\frac{\partial^2 u}{\partial x_i\,\partial x_k}\delta x_i\,dx_k\right]\Bigg|_{\substack{\delta x_1 = dx_1 \\ \cdots \\ \delta x_m = dx_m}} = \sum_{k=1}^{m}\sum_{i=1}^{m}\frac{\partial^2 u}{\partial x_i\,\partial x_k}dx_i dx_k. \quad (14.43)$$

(We have also made use of the fact that for a twice differentiable function mixed derivatives of the second order do not depend on the order of differentiation.)

Thus we find that when the arguments $x_1, x_2, \ldots, x_m$ are independent variables, the following representation is valid for the second differential of the function $u = f(x_1, x_2, \ldots, x_m)$ twice differentiable at a given point:

$$d^2u = \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{\partial^2 u}{\partial x_i \partial x_k} dx_i dx_k. \tag{14.44}$$

Remark 1. The function of m variables $t_1, t_2, \ldots, t_m$ of the form $\Phi = \sum_{i=1}^{m} \sum_{k=1}^{m} a_{ik} t_i t_k$, where a_{ik} are constant real numbers, is called the **quadratic form** of the variables $t_1, t_2, \ldots, t_m$ and the numbers a_{ik} are its coefficients.

The quadratic form is said to be symmetric if its coefficients comply with the condition $a_{ik} = a_{ki}$ (for all $i = 1, 2, \ldots, m$; $k = 1, 2, \ldots, m$).

Expression (14.44) we have obtained admits of an assertion that the arguments $x_1, x_2, \ldots, x_m$ being independent variables, the second differential of the function $u = f(x_1, x_2, \ldots, x_m)$ twice differentiable at the given point M is a symmetric* quadratic form of the variables $dx_1, dx_2, \ldots, dx_m$ whose coefficients are equal to the respective partial derivatives of the second order of the function $u = f(x_1, x_2, \ldots, x_m)$ taken at the given point M.

Make note of the fact that expression (14.44) for the second-order differential we have received can be rewritten in another form, using the conventional symbol

$$d = dx_1 \frac{\partial}{\partial x_1} + dx_2 \frac{\partial}{\partial x_2} + \ldots + dx_m \frac{\partial}{\partial x_m}. \tag{14.45}$$

With the aid of this symbol, expression (14.44) can be rewritten as

$$d^2u = \left(dx_1 \frac{\partial}{\partial x_1} + dx_2 \frac{\partial}{\partial x_2} + \ldots + dx_m \frac{\partial}{\partial x_m} \right)^2 u. \tag{14.46}$$

It can be easily verified by induction that when the arguments $x_1, x_2, \ldots, x_m$ of the function $u = f(x_1, x_2, \ldots, x_m)$ n times differentiable at the given point $M(x_1, x_2, \ldots, x_m)$ are independent variables, the following representation is valid for the nth differ-

---

* The quadratic form is symmetric because of the equality $\frac{\partial^2 u}{\partial x_i \partial x_k}(M) = \frac{\partial^2 u}{\partial x_k \partial x_i}(M)$.

ential of that function:

$$d^n u = \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \dots \sum_{i_n=1}^{m} \frac{\partial^n u}{\partial x_{i_1} \partial x_{i_2} \dots \partial x_{i_n}} dx_{i_1} dx_{i_2} \dots dx_{i_n}.$$

Using the conventional symbol (14.45), this representation can be rewritten as

$$d^n u = \left(dx_1 \frac{\partial}{\partial x_1} + dx_2 \frac{\partial}{\partial x_2} + \dots + dx_m \frac{\partial}{\partial x_m} \right)^n u. \tag{14.47}$$

The representations for the second and the subsequent differentials have quite different form in the case when the arguments $x_1, x_2, \dots, x_m$ of the function $u = f(x_1, x_2, \dots, x_m)$ are the functions of certain independent variables $t_1, t_2, \dots, t_k$ differentiable a corresponding number of times.

Let us discuss this latter case and find the expression for the second differential of the function $u = f(x_1, x_2, \dots, x_m)$ twice differentiable at the given point $M(x_1, x_2, \dots, x_m)$ whose arguments $x_1, x_2, \dots, x_m$ are twice differentiable functions of some independent variables $t_1, t_2, \dots, t_k$.

Repeating the reasoning used for the chain of equations (14.43), we obtain this time the following equations:

$$d^2 u = \delta (du) \Big|_{\substack{\delta x_1 = dx_1 \\ \dots\dots \\ \delta x_m = dx_m}} = \sum_{k=1}^{m} \left\{ dx_k \delta \left(\frac{\partial u}{\partial x_k} \right) + \frac{\partial u}{\partial x_k} \delta (dx_k) \right\} \Big|_{\substack{\delta x_1 = dx_1 \\ \dots\dots \\ \delta x_m = dx_m}} =$$

$$= \sum_{k=1}^{m} \left\{ dx_k \sum_{i=1}^{m} \frac{\partial}{\partial x_i} \left(\frac{\partial u}{\partial x_k} \right) \delta x_i \right\} \Big|_{\substack{\delta x_1 = dx_1 \\ \dots\dots \\ \delta x_m = dx_m}} =$$

$$= \sum_{k=1}^{m} \left\{ \frac{\partial u}{\partial x_k} \delta (dx_k) \right\} \Big|_{\substack{\delta x_1 = \delta x_1 \\ \dots\dots \\ \delta x_m = dx_m}}.$$

It should be pointed out that by virtue of the definition of the second differential of the function $u = x_k$ (where k is any of the numbers $1, 2, \dots, m$) we have

$$[\delta (dx_k)] \Big|_{\substack{\delta x_1 = dx_1 \\ \dots\dots \\ \delta x_m = dx_m}} = d^2 x_k.$$

Taking this relation into account, we arrive at the following representation for the second differential:

$$d^2 u = \sum_{k=1}^{m} \sum_{i=1}^{m} \frac{\partial^2 u}{\partial x_i \partial x_k} dx_i dx_k + \sum_{k=1}^{m} \frac{\partial u}{\partial x_k} d^2 x_k$$

or (making use of notation (14.45))

$$d^2u = \left(dx_1 \frac{\partial}{\partial x_1} + dx_2 \frac{\partial}{\partial x_2} + \ldots + dx_m \frac{\partial}{\partial x_m}\right)^2 u + \\ + \left(\frac{\partial u}{\partial x_1} d^2x_1 + \frac{\partial u}{\partial x_2} d^2x_2 + \ldots + \frac{\partial u}{\partial x_m} d^2x_m\right). \quad (14.48)$$

Comparing representation (14.48) we have received with representation (14.46), we shall ascertain that (as distinct from the first differential) the second differential no longer possesses the property of the invariance of the form. All the more so, this goes for all the subsequent differentials.

Remark 2. We wish to indicate an important special case where the second and the subsequent differentials of the function $u = f(x_1, x_2, \ldots, x_m)$ of m variables do possess the invariance of the form and are specified by the same formula (14.47) that is used for the case of the independent variables $x_1, x_2, \ldots, x_m$.

We shall say that the variables $x_1, x_2, \ldots, x_m$ are **linear functions** of the independent variables $t_1, t_2, \ldots, t_k$ if they are specified by the equalities

$$x_i = a_{i0} + a_{i1}t_1 + a_{i2}t_2 + \ldots + a_{ik}t_k \quad (i = 1, 2, \ldots, m),$$

where $a_{i0}, a_{i1}, \ldots, a_{ik}$ denote some constants.

Note that *if the function $u = f(x_1, x_2, \ldots, x_m)$ is n times differentiable at the given point $M(x_1, x_2, \ldots, x_m)$ and its arguments $x_1, x_2, \ldots, x_m$ are linear functions of the independent variables $t_1, t_2, \ldots, t_k$, then the nth differential of the function $u = f(x_1, x_2, \ldots, x_m)$ is specified by the same formula* (14.47) *that is used in the case of the independent variables $x_1, x_2, \ldots, x_m$.*

To verify this, we shall pay attention to the fact that since $t_1, t_2, \ldots, t_k$ are **independent** variables, the nth differential of x_i, as the functions of the arguments $t_1, t_2, \ldots, t_k$, is specified by the equality of the form (14.47), to be more exact, by the equality

$$d^n x_i = \left(dt_1 \frac{\partial}{\partial t_1} + dt_2 \frac{\partial}{\partial t_2} + \ldots + dt_k \frac{\partial}{\partial t_k}\right)^n x_i.$$

But any partial derivative of the order higher than the first of the linear function x_i is equal to zero. Consequently, $d^2x_i = 0$, $d^3x_i = 0, \ldots, d^n x_i = 0$.

Equality $d^2x_i = 0$ (for all $i = 1, 2, \ldots, m$) and representation (14.48) admit of an inference that d^2u is specified by equality (14.46). By a complete analogy, making use of the relations $d^3x_i = 0, \ldots, d^n x_i = 0$, we can prove by induction that $d^3u, d^4u, \ldots, d^n u$ are specified by equality (14.47).

Remark 3. When carrying out calculations, it is sometimes necessary to interpret equality (14.47) and, with due account of the fact that the equality includes coinciding terms, write out all the distinct terms of the equality with the coefficients preceding them.

For that purpose we can use **Newton's polynomial** formula which has the form

$$(a_1 + a_2 + \ldots + a_m)^n = \sum_{\substack{\alpha_1+\alpha_2+\ldots+\alpha_m=0 \\ 0\leqslant\alpha_i\leqslant n}} \frac{(\alpha_1+\alpha_2+\ldots+\alpha_m)!}{\alpha_1!\,\alpha_2!\,\ldots\,\alpha_m!}\, a_1^{\alpha_1} a_2^{\alpha_2} \ldots a_m^{\alpha_m} \qquad (14.49)$$

(the summation on the right-hand side of this formula is performed over all the integral-valued indices $\alpha_1, \alpha_2, \ldots, \alpha_m$, each of which satisfies the inequalities $0 \leqslant \alpha_i \leqslant n$ under the condition that the sum of all these indices $\alpha_1 + \alpha_2 + \ldots + \alpha_m$ is equal to n).

Formula (14.49) can be easily established by induction. Indeed, at $m = 2$ and any natural n, this formula is obviously valid since it reduces to the well-known Newton's binomial formula.

Assume that the formula is valid for some number $m \geqslant 2$ and any natural n and make sure that in such a case it also holds true for the number $m + 1$ and any natural n.

Representing $(a_1 + a_2 + \ldots + a_m + a_{m+1})^n$ in the form $(a_1 + a_2 + \ldots + a_m + a_{m+1})^n = [(a_1 + a_2 + \ldots + a_m) + a_{m+1}]^n$, let us use Newton's binomial to compute the coefficient in $a_1^{\alpha_1} a_2^{\alpha_2} \ldots a_m^{\alpha_m} a_{m+1}^{\alpha_{m+1}}$. By virtue of the equality $\alpha_1 + \alpha_2 + \ldots + \alpha_{m+1} = n$, Newton's binomial formula and the assumptions concerning the validity of formula (14.49) for the number m and any natural n, this coefficient is equal to*

$$C_n^{\alpha_{m+1}} \frac{(\alpha_1+\alpha_2+\ldots+\alpha_m)!}{(\alpha_1)!\,(\alpha_2)!\,\ldots\,(\alpha_m)!} = \frac{(\alpha_1+\alpha_2+\ldots+\alpha_m+\alpha_{m+1})!}{(\alpha_{m+1})!\,(\alpha_1+\alpha_2+\ldots+\alpha_m)!} \times$$

$$\times \frac{(\alpha_1+\alpha_2+\ldots+\alpha_m)!}{(\alpha_1)!\,(\alpha_2)!\,\ldots\,(\alpha_m)!} = \frac{(\alpha_1+\alpha_2+\ldots+\alpha_m+\alpha_{m+1})!}{(\alpha_1)!\,(\alpha_2)!\,\ldots\,(\alpha_m)!\,(\alpha_{m+1})!}.$$

The expression we have obtained for the coefficient in $a_1^{\alpha_1} a_2^{\alpha_2} \ldots a_m^{\alpha_m} a_{m+1}^{\alpha_{m+1}}$ exactly coincides with the expression which will result from formula (14.49) if in this formula we substitute $m + 1$ for m.

We have completed the induction and proved formula (14.49).

Formula (14.49) enables us to rewrite expression (14.47) for the nth differential in the following form:

$$d^n u\,(M) = \sum_{\substack{\alpha_1+\alpha_2+\ldots+\alpha_m=n \\ 0<\alpha_i<n}} \times$$

$$\times \frac{\partial^n u}{\partial x_1^{\alpha_1}\,\partial x_2^{\alpha_2}\,\ldots\,\partial x_m^{\alpha_m}}(M)\,(dx_1)^{\alpha_1}\,(dx_2)^{\alpha_2}\ldots(dx_m)^{\alpha_m}.$$

14.5.3. Taylor's formula for the function of m variables with the remainder in Lagrange's form. We shall use the notation $a^k u|_M$ for the differential of the kth order of the function $u = f(x_1, x_2, \ldots, x_m)$ at point M. Let us prove the following theorem.

* We take into account that $C_n^k = \frac{n!}{k!\,(n-k)!}$ and, consequently,

$$C_n^{\alpha_{m+1}} = C_{\alpha_1+\alpha_2+\ldots+\alpha_{m+1}}^{\alpha_{m+1}} = \frac{(\alpha_1+\alpha_2+\ldots+\alpha_m+\alpha_{m+1})!}{(\alpha_{m+1})!\,(\alpha_1+\alpha_2+\ldots+\alpha_m)!}.$$

Theorem 14.15. *Assume that the function* $u = f(M) = f(x_1, x_2, \ldots, x_m)$ *is defined in some* ε*-neighbourhood* of point* $M_0(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$ *and is* $n+1$ *times differentiable in the indicated* ε*-neighbourhood. Then the total increment* $\Delta u = f(M) - f(M_0)$ *of that function at point* M_0 *for any point* M *from the* ε*-neighbourhood can be represented in the following form:*

$$\Delta u = du\Big|_{M_0} + \frac{1}{2!}d^2u\Big|_{M_0} + \ldots + \frac{1}{n!}d^nu\Big|_{M_0} + \frac{1}{(n+1)!}d^{n+1}u\Big|_N. \tag{14.50}$$

In this case N *is some point from the indicated* ε*-neighbourhood dependent, in the general case, on* $M(x_1, x_2, \ldots, x_m)$, *and the differentials* dx_i *of the variables* x_i *appearing in the expressions* $d^ku|_{M_0}$ *and* $d^{n+1}u|_N$ *are equal to* $\Delta x_i = x_i - \overset{\circ}{x}_i$. Equation (14.50) is known as Taylor's formula for the function $u = f(M)$ with the centre of expansion at point M_0.

Proof. To abbreviate the notation, we shall carry out the argument for the function $u = f(x, y)$ of two variables x and y. We shall first write, in a special form, Taylor's formula for the function $u = F(t)$ of one variable t $n+1$ times differentiable in some neighbourhood of point t_0. Recall that Taylor's formula with the centre of expansion at t_0 for the function $u = F(t)$ of one variable is of the following form (the remainder is taken in Lagrange's form):

$$F(t) = F(t_0) + F'(t_0)(t-t_0) + \frac{1}{2!}F^{(2)}(t_0)(t-t_0)^2 + \ldots$$

$$\ldots + \frac{1}{n!}F^{(n)}(t_0)(t-t_0)^n + \frac{1}{(n+1)!}F^{(n+1)}(t_0+\theta(t-t_0))(t-t_0)^{n-1},$$

$$0 < \theta < 1. \tag{14.51}$$

The argument t being an independent variable, the increment $\Delta t = t - t_0$ is the differential dt of the independent variable t. Therefore,

$$F^{(k)}(t_0)(t-t_0)^k = F^{(k)}(t_0)\,dt^k = d^kF(t_0) = d^ku\,|_{t_0},$$

and

$$F^{(n+1)}(t_0+\theta(t-t_0))(t-t_0)^{n+1} = d^{n+1}u\,|_{t_0+\theta(t-t_0)}. \tag{14.52}$$

If we designate the difference $F(t) - F(t_0)$ as Δu, then, in accordance with (14.52), Taylor's formula (14.51) can be written in the

* Instead of the ε-neighbourhood of point M_0 we can take the so-called *convex neighbourhood* of that point which is defined as a neighbourhood of point M_0 which contains together with all its points M the entire segment M_0M.

following special form:

$$\Delta u = du\Big|_{t_0} + \frac{1}{2!} d^2u\Big|_{t_0} + \ldots + \frac{1}{n!} d^nu\Big|_{t_0} + \\ + \frac{1}{(n+1)!} d^{(n+1)}u\Big|_{t_0+\theta(t-t_0)}. \tag{14.53}$$

Let us now consider an arbitrary point $M(x_0 + \Delta x, y_0 + \Delta y)$ in the ε-neighbourhood of point $M_0(x_0, y_0)$ and connect points M_0 and M by a straight line. It is evident that the coordinates x and y of the points of this straight line are the following *linear functions* of the new variable t:

$$x = x_0 + t\Delta x, \quad y = y_0 + t\Delta y. \tag{14.54}$$

In this case the coordinates of the points of the segment M_0M corresponding to the values of the variable t belong to the closed interval $[0, 1]$. Note that the value $t = 0$ corresponds to point M_0 and the value $t = 1$ to point M. Since by the hypothesis the function $u = f(x, y)$ of two variables x and y is $n + 1$ times differentiable in the neighbourhood of point M_0 in question, it follows from formulas (14.54) that on the line M_0M this function is a composite function of the variable t, $(n + 1)$ times differentiable at least for all the values of t from the interval $[0, 1]$. Let us denote this composite function by $F(t)$ and write for it Taylor's formula with the centre of expansion at point $t_0 = 0$ in the special form (14.53) with

$$\Delta u = F(1) - F(0) = f(M) - f(M_0).$$

Differentials of different orders appearing in formula (14.53) are the differentials of the composite function $u = f(x, y)$, where x and y are linear functions (14.54). According to the remark made in the previous subsection, under these conditions differentials of any order of the function $u = f(x, y)$ can be written in the form (14.47). Therefore,

$$\left.\begin{aligned} &d^ku\Big|_{t_0=0} = \left(\frac{\partial}{\partial x}dx + \frac{\partial}{\partial y}dy\right)^k u\Big|_{M_0(x_0, y_0)} = d^ku\,|_{M_0}, \\ &d^{n+1}u\Big|_{t_0+\theta(t-t_0)} = \left(\frac{\partial}{\partial x}dx + \frac{\partial}{\partial y}dy\right)^{n+1} u\Big|_{N(x_0+\theta\Delta x,\ y_0+\theta\Delta y)} = \\ &= d^{n+1}u\,|_N. \end{aligned}\right\} \tag{14.55}$$

The quantities dx and dy in formulas (14.55) can be found from relations (14.54) at $dt = \Delta t = 1 - 0 = 1$. Thus, in formulas (14.55)

$$dx = dt\,\Delta x = \Delta x \text{ and } dy = dt\,\Delta y = \Delta y. \tag{14.56}$$

Substituting $d^ku\,|_{t_0}$ and $d^{n+1}u\,|_{t_0+\theta(t-t_0)}$ from (14.55) into (14.53) and taking into account relations (14.56), we obtain Taylor's formula (14.50). The proof of the theorem is complete.

Written out in detail below is the expression of Taylor's formula (14.50) for function $u = f(x_1, x_2, \ldots, x_m)$:

$$f(x_1, x_2, \ldots, x_m) = f(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m) +$$
$$+ \sum_{k=1}^{n} \frac{1}{k!} \left[\frac{\partial}{\partial x_1}(x_1 - \overset{\circ}{x}_1) + \frac{\partial}{\partial x_2}(x_2 - \overset{\circ}{x}_2) + \ldots \right.$$
$$\left. \ldots + \frac{\partial}{\partial x_m}(x_m - \overset{\circ}{x}_m) \right]^k f(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m +$$
$$+ \frac{1}{(n+1)!} \left[\frac{\partial}{\partial x_1}(x_1 - \overset{\circ}{x}_1) + \frac{\partial}{\partial x_2}(x_2 - \overset{\circ}{x}_2) + \ldots \right.$$
$$\left. \ldots + \frac{\partial}{\partial x_m}(x_m - \overset{\circ}{x}_m) \right]^{n+1} f[\overset{\bullet}{x}_1 + \theta(x_1 - \overset{\bullet}{x}_1),$$
$$\overset{\bullet}{x}_2 + \theta(x_2 - \overset{\circ}{x}_2), \ldots, \overset{\circ}{x}_m + \theta(x_m - \overset{\bullet}{x}_m)]. \qquad (14.57)$$

14.5.4. Taylor's formula with the remainder in Peano's form. ***Theorem 14.15.**** *Assume that $n \geqslant 1$ is an integer, the function $u = f(M) = f(x_1, x_2, \ldots, x_m)$ is defined and $n-1$ times differentiable in the ε-neighbourhood of point $M_0(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$ and n times differentiable at point M_0 itself.* Then at any point M from the indicated ε-neighbourhood of M_0 the following formula holds true:*

$$f(M) = f(M_0) + \frac{1}{1!} du\Big|_{M_0} + \frac{1}{2!} d^2u\Big|_{M_0} + \ldots + \frac{1}{n!} d^nu\Big|_{M_0} + o(\rho^n), \qquad (14.58)$$

in which ρ denotes the distance $\rho(M_0M)$ and $o(\rho^n)$ denotes an infinitesimal function, at $\rho \to 0$ (or at $M \to M_0$), of the higher order of smallness than ρ^n.

Equation (14.58) is called *Taylor's formula* (with centre at point M_0) with the remainder *in Peano's form.*

Remark. Written out in more detail Taylor's formula (14.58) assumes the form

$$f(x_1, x_2, \ldots, x_m) = f(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m) +$$
$$+ \sum_{k=1}^{n} \frac{1}{k!} \left[(x_1 - \overset{\bullet}{x}_1) \frac{\partial}{\partial x_1} + \ldots + (x_m - \overset{\bullet}{x}_m) \frac{\partial}{\partial x_m} \right]^k \times$$
$$\times f(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\bullet}{x}_m) + o(\rho^n). \qquad (14.59)$$

* At $n = 1$, a requirement must be put that the function $u = f(x_1, x_2, \ldots, x_m)$ should be given only in the ε-neighbourhood of point M_0 and differentiable at point M_0 itself.

Let us pay attention to the fact that the right-hand side of (14.59) includes the sum of the polynomial of degree n of the variables $x_1, x_2, \ldots, x_m$ and the remainder $o(\rho^n)$.

Let us denote by $R_{n+1}(M)$ the difference between $f(M)$ and the indicated polynomial, i.e. put

$$R_{n+1}(M) = f(M) - f(M_0) - \sum_{k=1}^{n} \frac{1}{k!}\left[(x_1 - \overset{\circ}{x}_1)\frac{\partial}{\partial x_1} + \ldots\right.$$

$$\left. \ldots + (x_m - \overset{\circ}{x}_m)\frac{\partial}{\partial x_m}\right]^k f(M_0). \qquad (14.60)$$

Theorem 14.15* will be proved if we determine that $R_{n+1}(M) = o(\rho^n)$ when the conditions of the theorem are satisfied.

We shall precede the proof of Theorem 14.15* by two lemmas.

Lemma 1. *If the function* $f(M) = f(x_1, x_2, \ldots, x_m)$ *is* n *times differentiable at point* $M_0(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$, *then both the function* $R_{n+1}(M)$ *itself, specified by equality* (14.60), *and all its partial derivatives with respect to any variables* $x_1, x_2, \ldots, x_m$ *up to order* n *inclusive, vanish at point* M_0.

Proof. At $n = 1$ function (14.60) assumes the form

$$R_2(M) = f(M) - f(M_0) - (x_1 - \overset{\circ}{x}_1)\frac{\partial f}{\partial x_1}(M_0) - \ldots$$

$$\ldots - (x_m - \overset{\circ}{x}_m)\frac{\partial f}{\partial x_m}(M_0),$$

while the equalities $R_2(M_0) = 0$, $\frac{\partial R_2}{\partial x_i}(M_0) = 0$ (for all $i = 1, 2, \ldots, m$) are easy to verify.

To use the induction method, let us assume that the lemma holds true for some number $n \geqslant 1$ and prove that in that case it is valid for the number $n + 1$ as well.

Assume that the function $f(M)$ is $n + 1$ times differentiable at point M_0 and that

$$R_{n+2}(M) = f(M) - f(M_0) - \sum_{k=1}^{n+1} \frac{1}{k!}\left[(x_1 - \overset{\circ}{x}_1)\frac{\partial}{\partial x_1} + \ldots\right.$$

$$\left. \ldots + (x_m - \overset{\circ}{x}_m)\frac{\partial}{\partial x_m}\right]^k f(M_0). \qquad (14.61)$$

The equality $R_{n+2}(M_0) = 0$ is easy to verify (it is sufficient to take into account that the expression in the parentheses $(x_i - \overset{\circ}{x}_i)$ in (14.61) vanishes at point M_0).

It remains to prove that for any $i = 1, 2, \ldots, m$ the function $\frac{\partial R_{n+2}}{\partial x_i}(M)$ itself and all its partial derivatives up to order n

inclusive vanish at point M_0, and for that purpose, by virtue of the assumption we have made concerning the validity of the lemma for the number n, it is sufficient to prove that the function $\frac{\partial R_{n+2}}{\partial x_i}(M)$ is specified by the equality of the form of (14.60), to be more exact, by the equality

$$\frac{\partial R_{n+2}}{\partial x_i}(M) = \frac{\partial f}{\partial x_i}(M) - \frac{\partial f}{\partial x_i}(M_0) -$$

$$- \sum_{k=1}^{n} \frac{1}{k!} \left[(x_1 - \overset{\circ}{x}_1) \frac{\partial}{\partial x_1} + \ldots + (x_m - \overset{\circ}{x}_m) \frac{\partial}{\partial x_m} \right]^k \frac{\partial f}{\partial x_i}(M_0). \quad (14.62)$$

Since the independent variables x_i $(i = 1, 2, \ldots, m)$ are involved in the expression for $R_{n+2}(M)$ symmetrically, it is sufficient to prove equality (14.62) for $i = 1$, that is, to prove the equality

$$\frac{\partial R_{n+2}}{\partial x_1}(M) = \frac{\partial f}{\partial x_1}(M) - \frac{\partial f}{\partial x_1}(M_0) -$$

$$- \sum_{k=1}^{n} \frac{1}{k!} \left[(x_1 - \overset{\circ}{x}_1) \frac{\partial}{\partial x_1} + \ldots + (x_m - \overset{\circ}{x}_m) \frac{\partial}{\partial x_m} \right]^k \frac{\partial f}{\partial x_1}(M_0). \quad (14.63)$$

It is evident from (14.61) that to prove (14.63) it is sufficient to make sure that for every number $k = 1, 2, \ldots, n+1$ at fixed $x_2, x_3, \ldots, x_m$

$$\frac{d}{dx_1} \left[(x_1 - \overset{\circ}{x}_1) \frac{\partial}{\partial x_1} + (x_2 - \overset{\circ}{x}_2) \frac{\partial}{\partial x_2} + \ldots + (x_m + \overset{\circ}{x}_m) \frac{\partial}{\partial x_m} \right]^k f(M_0) =$$

$$= k \left[(x_1 - \overset{\circ}{x}_1) \frac{\partial}{\partial x_1} + (x_2 - \overset{\circ}{x}_2) \frac{\partial}{\partial x_2} + \ldots \right.$$

$$\left. \ldots + (x_m - \overset{\circ}{x}_m) \frac{\partial}{\partial x_m} \right]^{k-1} \frac{\partial f}{\partial x_1}(M_0). \quad (14.64)$$

Since we differentiate with respect to x_1, with the variables $x_2, x_3, \ldots, x_m$ being fixed, the quantity

$$D = (x_2 - \overset{\circ}{x}_2) \frac{\partial}{\partial x_2} + \ldots + (x_m - \overset{\circ}{x}_m) \frac{\partial}{\partial x_m},$$

in differentiating with respect to x_1, can be regarded as constant. It should be added that inasmuch as the symbols $\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_m}$ are used for the formation of partial derivatives of the function f at a *fixed* point M_0, these symbols should also be regarded as constant when differentiating with respect to x_1.

By virtue of the aforesaid, to prove equality (14.64) it is sufficient to verify the validity of the equality

$$\frac{d}{dx_1}\left[(x_1-\overset{\bullet}{x}_1)\frac{\partial}{\partial x_1}+D\right]^k = k\left[(x_1-\overset{\bullet}{x}_1)\frac{\partial}{\partial x_1}+D\right]^{k-1}\frac{\partial}{\partial x_1}. \qquad (14.65)$$

Differentiating the function $\left[(x_1-\overset{\bullet}{x}_1)\frac{\partial}{\partial x_1}+D\right]^k$ with respect to x_1 as a composite function and recalling the independence of D and $\frac{\partial}{\partial x_1}$ of x_1, we obtain equality (14.65). The induction is complete.

We have proved Lemma 1.

Lemma 2. *Assume that $R(M) = R(x_1, x_2, \ldots, x_m)$ is an arbitrary function satisfying two requirements:*

(1) *$R(M)$ is n times differentiable at point $M_0(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$,*

(2) *the function $R(M)$ itself and all its partial derivatives with respect to any of the variables $x_1, x_2, \ldots, x_m$ up to order n inclusive vanish at the indicated point M_0. Then for the function $R(M)$ there holds an estimation*

$$R(M) = o(\rho^n), \qquad (14.66)$$

in which ρ denotes the distance $\rho(M_0M)$ between points M_0 and M.

Proof. For $n = 1$ the proof of the lemma follows from the differentiation condition* of the function $R(M)$ at point M_0, which has the form

$$R(M) - R(M_0) = \sum_{k=1}^{m} \frac{\partial R}{\partial x_k}(M_0)(x_k - \overset{\bullet}{x}_k) + o(\rho).$$

Taking into account that $R(M_0) = 0$, $\frac{\partial R}{\partial x_k}(M_0) = 0$ for all $k = 1, 2, \ldots, m$ we find that $R(M) = o(\rho^n)$.

To carry out inductive reasoning, let us assume that Lemma 2 is valid for some number $n \geqslant 1$ and prove that then it is also valid for the number $n + 1$.

Let the function $R(M)$ satisfy the two requirements of Lemma 2 for the number $n+1$. Then, evidently, any first-order partial derivative of that function $\frac{\partial R}{\partial x_k}(M)$ $(k = 1, 2, \ldots, m)$ will satisfy the two requirements of Lemma 2 for the number n, and therefore (by virtue of the supposition we have made concerning the validity of Lemma 2 for the number n), the estimation

$$\frac{\partial R}{\partial x_k}(M) = o(\rho^n) \qquad (14.66^*)$$

holds true.

* See relation (14.16) in 14.4.2.

Let us note that since $n \geqslant 1$, it follows that $n+1 \geqslant 2$ and the function $R(M)$ satisfying the two requirements of Lemma 2 for the number $n+1$ is at any rate once differentiable in the neighbourhood of point M_0. Therefore, the conditions of Theorem 14.15 are complied with for that function for the number $n=0$. In accordance with the indicated theorem, for any point M from a sufficiently small ε-neighbourhood of point M_0 on the segment M_0M there is a point N such that there holds a formula

$$R(M)=R(M_0)+\frac{1}{1!}\sum_{k=1}^{m}(x_k-\overset{\circ}{x}_k)\frac{\partial R}{\partial x_k}(N). \tag{14.67}$$

Let us now take notice of the fact that since point N lies between points M_0 and M and ρ is the distance between points M_0 and M, it follows that $\rho(N, M_0) \leqslant \rho$, and consequently (14.66*) yields $\frac{\partial R}{\partial x_k}(N)=o(\rho^n)$.

Substituting the last estimation into (14.67) and taking into account that $R(M_0)=0$, we find that

$$R(M)=o(\rho^n)\sum_{k=1}^{m}|x_k-\overset{\circ}{x}_k|.$$

Since $|x_k-\overset{\circ}{x}_k| \leqslant \sqrt{\sum_{i=1}^{m}(x_i-\overset{\circ}{x}_i)^2}=\rho$, we finally receive $R(M)=o(\rho^n)$.

We have completed the induction and proved Lemma 2.

Proof of Theorem 14.15* can be easily carried out with the aid of Lemmas 1 and 2.

Indeed, as was mentioned above, to prove Theorem 14.15* it is sufficient to determine that when the conditions of the theorem are satisfied, the following evaluation holds for the function (14.60):

$$R_{n+1}(M)=o(\rho^n).$$

By virtue of Lemma 1 the function (14.60) itself and all its partial derivatives with respect to any of the variables $x_1, x_2, \ldots, x_m$ up to order n inclusive vanish at point M_0. But then, in accordance with Lemma 2, the evaluation $R_{n+1}(M)=o(\rho^n)$ is valid for the function (14.60).

We have proved Theorem 14.15*.

14.6. A LOCAL EXTREMUM OF A FUNCTION OF m VARIABLES

14.6.1. The concept of an extremum of a function of m variables. Necessary conditions of an extremum. Suppose that a function $u=f(M)=f(x_1, x_2, \ldots, x_m)$ of m variables is defined in some neighbourhood of point $M_0(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$ of the space E^m.

Definition 1. *We shall say that the function $u = f(M)$ possesses a* **local maximum (local minimum)** *at point M_0 if there is a δ-neighbourhood of point M_0 throughout which the value $f(M_0)$ is the greatest (the least) of all the values $f(M)$ of that function.*

Definition 2. *We shall say that the function $u = f(M)$ possesses a* **local extremum** *at point M_0 if it possesses a local maximum or a local minimum at that point.*

Let us establish the necessary conditions of a local extremum of the function $u = f(M)$ possessing at the given point M_0 partial derivatives of the first order with respect to all the variables.

Let us prove the following **assertion**: *if the function $u = f(M) = f(x_1, x_2, \ldots, x_m)$ possesses at point $M_0(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$ partial derivatives of the first order with respect to all the variables $x_1, x_2, \ldots, x_m$ and possesses a local extremum at that point, then all the partial derivatives of the first order vanish at point M_0, that is, the following equalities hold true:*

$$\frac{\partial u}{\partial x_1}(M_0) = 0, \quad \frac{\partial u}{\partial x_2}(M_0) = 0, \ldots, \frac{\partial u}{\partial x_m}(M_0) = 0. \tag{14.68}$$

Proof. We shall establish the validity of the first equality (14.68). We fix the arguments $x_2, x_3, \ldots, x_m$ of the function $u = f(x_1, x_2, \ldots, x_m)$ putting them equal to the respective coordinates of point M_0, that is, putting $x_2 = \overset{\circ}{x}_2$, $x_3 = \overset{\circ}{x}_3$, $\ldots$, $x_m = \overset{\circ}{x}_m$. As a result we obtain a function $u = f(x_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$ of one variable x_1. The derivative of this function of one variable at point $x_1 = \overset{\circ}{x}_1$ coincides with the partial derivative $\frac{du}{dx_1}(M_0)$.

Since the function $u = f(M)$ of m variables possesses a local extremum at point M_0, the indicated function $u = f(x_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$ of one variable possesses a local extremum at point $x_1 = \overset{\circ}{x}_1$, and, therefore (by virtue of the results obtained in 8.7.2), the derivative of this function of one variable at point $x_1 = \overset{\circ}{x}_1$, coinciding with the partial derivative $\frac{du}{dx_1}(M_0)$, is equal to zero.

We have proved the first equality (14.68); the rest of the equalities (14.68) can be proved by analogy.

We wish to emphasize that equalities (14.68) (i.e. vanishing of all the first-order derivatives at the given point M_0) are only the necessary but not the sufficient conditions of a local extremum of the function $u = f(M)$ at point M_0.

For example, both partial derivatives $\partial u/\partial x$ and $\partial u/\partial y$ of the function $u = x \cdot y$ of two variables vanish at point $M_0(0, 0)$, but the indicated function does not possess an extremum at that point $M_0(0, 0)$ since that function $u = x \cdot y$ is equal to zero at the point

$M_0(0, 0)$ itself, and in the arbitrarily small δ-neighbourhood of that point it can assume positive as well as negative values.

The points at which all the first-order partial derivatives of the function $u = f(M)$ vanish are called **points of a possible extremum** of that function.

At each point of its possible extremum the function $u = f(M)$ can possess a local extremum, but the existence of this local extremum can be established only with the aid of the sufficient conditions of a local extremum which will be elucidated in the next subsection.

The assertion we have proved yields another form of the necessary conditions of a local extremum:

if the function $u = f(M)$ is differentiable at point M_0 and possesses a local extremum at that point, then the differential $du|_{M_0}$ of the function at point M_0 is identically equal to zero with respect to the differentials of the independent variables $dx_1, dx_2, \ldots, dx_m$.

Indeed, since

$$du|_{M_0} = \frac{\partial u}{\partial x_1}(M_0)\,dx_1 + \frac{\partial u}{\partial x_2}(M_0)\,dx_2 + \ldots + \frac{\partial u}{\partial x_m}(M_0)\,dx_m,$$

it follows from equalities (14.68) that *for any* $dx_1, dx_2, \ldots, dx_m$ the equality $du|_{M_0} = 0$ holds true.

14.6.2. Sufficient conditions of a local extremum of a function of m variables. In formulating the sufficient conditions for a local extremum of the function $u = f(M)$ of m variables a significant role will be played by the second differential of that function at the point M_0 in question.

We have made sure in 14.5.2 that for the case where the arguments $x_1, x_2, \ldots, x_m$ of the twice differentiable function $u = f(x_1, x_2, \ldots, x_m)$ are either independent variables or linear functions of some independent variables, the second differential of this function at the given point M_0 is the quadratic form relative to the differentials of the arguments $dx_1, dx_2, \ldots, dx_m$ of the following form:

$$d^2u|_{M_0} = \sum_{i=1}^{m}\sum_{k=1}^{m} a_{ik}\,dx_i\,dx_k, \tag{14.69}$$

where

$$a_{ik} = a_{ki} = \frac{\partial^2 u}{\partial x_i\,\partial x_k}(M_0). \tag{14.70}$$

To formulate the sufficient conditions for a local extremum, we shall need some information from the theory of quadratic forms which we present below for the convenience of the reader*.

* All the definitions and statements given here can be found in any course of linear algebra.

The quadratic form with respect to the variables $h_1, h_2, \ldots, h_m$

$$\Phi(h_1, h_2, \ldots, h_m) = \sum_{i=1}^{m} \sum_{k=1}^{m} a_{ik} h_i h_k \tag{14.71}$$

is said to be **positive definite (negative definite)** if for any values $h_1, h_2, \ldots, h_m$ which are simultaneously nonzero, this form assumes strictly positive (strictly negative) values.

Quadratic form (14.71) is called *definite* if it is either positive definite or negative definite.

Quadratic form (14.71) is called *alternating* if it can assume both strictly positive and strictly negative values.

Quadratic form (14.71) is called *semi-definite* if it assumes either only nonnegative or only nonpositive values and vanishes for the values $h_1, h_2, \ldots, h_m$ which are simultaneously nonzero.

Let us formulate the so-called Sylvester* theorem on the definiteness of a quadratic form.

We shall say that the following matrix is a matrix of the quadratic form (14.71):

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mm} \end{pmatrix}. \tag{14.72}$$

If all the elements of the matrix A satisfy the condition $a_{ik} = a_{ki}$ $(i = 1, 2, \ldots, m;\ k = 1, 2, \ldots, m)$, then the matrix is called *symmetric*.

We shall use the term *principal minors* of the symmetric matrix (14.72) for the following determinants:

$$A_1 = a_{11},\quad A_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix},\quad A_3 = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}, \ldots$$

$$\ldots,\ A_m = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mm} \end{vmatrix}.$$

Sylvester's theorem is usually formulated as the following two assertions:

1°. *For the quadratic form* (14.71) *with the symmetric matrix* (14.72) *to be positive definite, it is necessary and sufficient that the principal minors of the matrix* (14.72) *should all be positive, i.e. that the follow-*

* James Joseph Sylvester (1814-1897), an English mathematician.

ing inequalities should hold true:

$$A_1 > 0,\ A_2 > 0,\ \ldots,\ A_m > 0.$$

2°. *For the quadratic form* (14.71) *with the symmetric matrix* (14.72) *to be negative definite, it is necessary and sufficient that the signs of the principal minors of the matrix* (14.72) *should alternate, with the sign of* A_1 *being negative, i.e. that the following inequalities should hold true*:

$$A_1 < 0,\ A_2 > 0,\ A_3 < 0,\ A_4 > 0,\ \ldots$$

Now everything is prepared for the formulation and proof of the theorem establishing the sufficient conditions of a local extremum.

Theorem 14.16. *Let the function* $u = f(M) = f(x_1, x_2, \ldots, x_m)$ *of* m *variables be once differentiable in some neighbourhood of point* $M_0(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$ *and twice differentiable at point* M_0 *itself. Assume, in addition, that point* M_0 *is a point of a possible extremum of the function* $u = f(M)$, *i.e.* $du\,|_{M_0} = 0$. *Then, if the second differential* (14.69)-(14.70) *is a positive definite* (*negative definite*) *quadratic form of the variables* $dx_1, dx_2, \ldots, dx_m$, *the function* $u = f(M)$ *possesses a local minimum* (*local maximum*) *at point* M_0. *Now if the second differential* (14.69)-(14.70) *is an alternating quadratic form, then the function* $u = f(M)$ *does not possess a local extremum at point* M.

Proof. We shall first prove the first part of the theorem, assuming for definiteness that the second differential (14.69)-(14.70) is a positive definite quadratic form of the variables $dx_1, dx_2, \ldots, dx_m$. Let us prove that in this case the function $u = f(M)$ possesses a local minimum at point M_0.

Let us expand the function $u = f(M)$ in the neighbourhood of point M_0 by Taylor's formula with the remainder in Peano's form, taking in this formula $n = 2$*. The result will be

$$f(M) = f(M_0) = du\,|_{M_0} + \frac{1}{2!}\,d^2u\,|_{M_0} + o(\rho^2), \tag{14.73}$$

with the differentials dx_k of the variables x_k in equality (14.73), entering into the expressions for $du\,|_{M_0}$ and $d^2u|_{M_0}$, being equal to the respective increments $(x_k - \overset{\circ}{x}_k)$ of those variables, and the quantity ρ being equal to

$$\rho = \sqrt{(dx_1)^2 + (dx_2)^2 + \ldots + (dx_m)^2} = \\ = \sqrt{(x_1 - \overset{\circ}{x}_1)^2 + (x_2 - \overset{\circ}{x}_2)^2 + \ldots + (x_m - \overset{\circ}{x}_m)^2}. \tag{14.74}$$

* All the conditions of Theorem 14.15* (see 14.5.4) are fulfilled for the function $u = f(M)$ for $n = 2$.

By the hypothesis, point M_0 is a point of a possible extremum. Therefore, on the basis of the results obtained in the previous subsection $du\,|_{M_0} = 0$. Taking this equality into account and putting $dx_k = x_k - \overset{\circ}{x}_k$ in expressions (14.69)-(14.70) for the second differential, we shall impart to Taylor's formula (14.73) the following form:

$$f(M) - f(M_0) = \frac{1}{2} \sum_{i=1}^{m} \sum_{k=1}^{m} a_{ik} (x_i - \overset{\circ}{x}_i)(x_k - \overset{\circ}{x}_k) + o(\rho^2). \qquad (14.75)$$

It suffices to prove that for all sufficiently small ρ the right-hand side of (14.75) is positive (and this will mean that in a sufficiently small neighbourhood of point M_0 the difference $f(M) - f(M_0)$ is positive, i.e. the function $u = f(M)$ possesses a local minimum at point M_0).

Let us put $h_i = \dfrac{x_i - \overset{\circ}{x}_i}{\rho}$, where $i = 1, 2, \ldots, m$. Then the expression (14.74) for ρ yields the following relations:

$$|h_i| \leqslant 1, \quad h_1^2 + h_2^2 + \ldots + h_m^2 = 1. \qquad (14.76)$$

With the aid of the notation we have introduced, equality (14.75) can be rewritten as follows:

$$f(M) - f(M_0) = \frac{\rho^2}{2} \sum_{i=1}^{m} \sum_{k=1}^{m} a_{ik} h_i h_k + o(\rho^2). \qquad (14.75^*)$$

The ratio $o(\rho^2)/\rho^2$ is an infinitely small function at $\rho \to 0$ (or at $M \to M_0$), which will be denoted by $\alpha(\rho)$. The introduction of this function makes it possible to write the equality $o(\rho^2) = \rho^2 \cdot \alpha(\rho)$ which will help us to impart to relation (14.75*) the form

$$f(M) - f(M_0) = \rho^2 \left[\frac{1}{2} \sum_{i=1}^{m} \sum_{k=1}^{m} a_{ik} h_i h_k + \alpha(\rho) \right]. \qquad (14.75^{**})$$

It is now easy to prove that the right-hand side of (14.75**) is positive for all sufficiently small ρ. The quadratic form $\Phi = \sum_{i=1}^{m} \sum_{k=1}^{m} a_{ik} h_i h_k$ is a function defined and continuous on the surface of unit sphere (14.76) which is a closed and bounded set. In accordance with Weierstrass's second theorem (see Theorem 14.7 in 14.3.2), this function reaches its greatest lower bound μ on this set, and it follows from the positive definiteness of the quadratic form (14.71) and from the fact that $h_1, h_2, \ldots, h_m$ satisfying relation (14.76) are not simultaneously zero that the indicated greatest lower bound μ is *strictly positive*.

Since the function $\alpha(\rho)$ infinitesimal at $\rho \to 0$ satisfies the inequality $|\alpha(\rho)| < \mu$ for all sufficiently small ρ the entire right-hand side of (14.75) is positive at all sufficiently small ρ, i.e. at all M sufficiently close to M_0.

And this precisely means that the function $u = f(M)$ possesses a local minimum at point M_0.

It can be proved by a complete analogy that when the second differential (14.69)-(14.70) is a negative definite quadratic form, the function $u = f(M)$ possesses a local maximum at point M_0.

We shall now prove the second part of the theorem, that is, we shall prove that in the case where the second differential (14.69)-(14.70) is an alternating quadratic form, the function $u = f(M)$ does not possess a local extremum at point M_0.

Let us establish, first of all, the following auxiliary property of the alternating quadratic form (14.71).

If the quadratic form $\Phi(h_1, h_2, \ldots, h_m)$ is alternating, then there are two totalities of the variables $(h_1', h_2', \ldots, h_m')$ and $(h_1'', h_2'', \ldots, h_m'')$ such that

$$(h_1')^2 + (h_2')^2 + \ldots + (h_m')^2 = 1, \quad (h_1'')^2 + (h_2'')^2 + \ldots + (h_m'')^2 = 1, \tag{14.77}$$

with

$$\Phi(h_1', h_2', \ldots, h_m') > 0, \quad \Phi(h_1'', h_2'', \ldots, h_m'') < 0. \tag{14.78}$$

Indeed, in accordance with the definition of an alternating quadratic form, there are two totalities of the arguments $(t_1', t_2', \ldots, t_m')$ and $(t_1'', t_2'', \ldots, t_m'')$ consisting of numbers which are not simultaneously zero and such that

$$\Phi(t_1', t_2', \ldots, t_m') > 0, \quad \Phi(t_1'', t_2'', \ldots, t_m'') < 0. \tag{14.79}$$

Putting

$$h_i' = \frac{t_i'}{\sqrt{(t_1')^2 + (t_2')^2 + \ldots + (t_m')^2}}, \qquad h_i'' = \frac{t_i''}{\sqrt{(t_1'')^2 + (t_2'')^2 + \ldots + (t_m'')^2}} \tag{14.80}$$

and taking into consideration that it immediately follows from the definition (14.71) of the quadratic form that

$$\Phi(h_1', h_2' \ldots, h_m') = \frac{1}{(t_1')^2 + (t_2')^2 + \ldots + (t_m')^2} \Phi(t_1', t_2', \ldots, t_m'),$$

$$\Phi(h_1'', h_2'', \ldots, h_m'') = \frac{1}{(t_1'')^2 + (t_2'')^2 + \ldots + (t_m'')^2} \Phi(t_1'', t_2'', \ldots, t_m''),$$

we obtain (by virtue of (14.79)) inequalities (14.78), with inequalities (14.77) immediately following from relations (14.80).

We have proved the auxiliary property of an alternating quadratic form.

Let us now proceed with the proof of the second part of the theorem. We shall fix two totalities of the variables $(h'_1, h'_2, \ldots, h'_m)$ and $(h''_1, h''_2, \ldots, h''_m)$ satisfying relations (14.77) and (14.78) and prove that for any $\rho > 0$ there are two points $M'(x'_1, x'_2, \ldots, x'_m)$ and $M''(x''_1, x''_2, \ldots, x''_m)$ of the space E^m such that $\rho(M', M_0) = \rho(M'', M_0) = \rho$, with

$$\frac{x'_i - \overset{\circ}{x}_i}{\rho} = h'_i, \quad \frac{x''_i - \overset{\circ}{x}_i}{\rho} = h''_i \text{ for all } i = 1, 2, \ldots, m. \tag{14.81}$$

In fact, putting

$$x'_i = \overset{\circ}{x}_i + \rho h'_i, \quad x''_i = \overset{\circ}{x}_i + \rho h''_i$$

for any $\rho > 0$ and for every number i $(i = 1, 2, \ldots, m)$, we shall comply with relations (14.81). In this case, by virtue of equalities (14.77), the following equalities will hold true:

$$\rho(M', M_0) = \sqrt{(x'_1 - \overset{\circ}{x}_1)^2 + (x'_2 - \overset{\circ}{x}_2)^2 + \ldots + (x'_m - \overset{\circ}{x}_m)^2} =$$
$$= \rho \sqrt{(h'_1)^2 + (h'_2)^2 + \ldots + (h'_m)^2} = \rho,$$
$$\rho(M'', M_0) = \sqrt{(x''_1 - \overset{\circ}{x}_1)^2 + (x''_2 - \overset{\circ}{x}_2)^2 + \ldots + (x''_m - \overset{\circ}{x}_m)^2} =$$
$$= \rho \sqrt{(h''_1)^2 + (h''_2)^2 + \ldots + (h''_m)^2} = \rho.$$

It is now easy to verify that for the case of the second differential (14.69)-(14.70) being an alternating quadratic form, the function $u = f(M)$ does not have an extremum at point M_0.

Writing out the expansion of the function $u = f(M)$ in the neighbourhood of point M_0 by Taylor's formula with the remainder in Peano's form and taking this expansion at the indicated points M' and M'', we obtain, instead of (14.75), the following two expansions:

$$f(M') - f(M_0) = \frac{1}{2} \sum_{i=1}^{m} \sum_{k=1}^{m} a_{ik} (x'_i - \overset{\circ}{x}_i)(x'_k - \overset{\circ}{x}_k) + o(\rho^2), \tag{14.82}$$

$$f(M'') - f(M_0) = \frac{1}{2} \sum_{i=1}^{m} \sum_{k=1}^{m} a_{ik} (x''_i - \overset{\circ}{x}_i)(x''_k - \overset{\circ}{x}_k) + o(\rho^2), \tag{14.83}$$

which are valid for all sufficiently small $\rho > 0$.

Substituting the values $(x'_i - \overset{\circ}{x}_i)$ and $(x''_i - \overset{\circ}{x}_i)$ from equalities (14.81) into these expansions and taking into account that $o(\rho^2) = \rho^2 \alpha(\rho)$, where $\alpha(\rho) \to 0$ at $\rho \to 0$, we shall impart to expansions

(14.82) and (14.83) the following form:

$$f(M')-f(M_0)=\rho^2\left[\frac{1}{2}\sum_{i=1}^{m}\sum_{k=1}^{m}a_{ik}h_i'h_k'+\alpha(\rho)\right],$$

$$f(M'')-f(M_0)=\rho^2\left[\frac{1}{2}\sum_{i=1}^{m}\sum_{k=1}^{m}a_{ik}h_i''h_k''+\alpha(\rho)\right].$$

The last two relations can be rewritten as follows:

$$f(M')-f(M_0)=\rho^2\left[\frac{1}{2}\Phi(h_1', h_2', \ldots, h_m')+\alpha(\rho)\right], \qquad (14.82^*)$$

$$f(M'')-f(M_0)=\rho^2\left[\frac{1}{2}\Phi(h_1'', h_2'', \ldots, h_m'')+\alpha(\rho)\right]. \qquad (14.83^*)$$

Taking into consideration relations (14.78) and the fact that the quantities $\Phi(h_1', h_2', \ldots, h_m') > 0$ and $\Phi(h_1'', h_2'', \ldots, h_m'') < 0$ do not depend on ρ and recalling that $\rho = \rho(M', M_0) = \rho(M'', M_0)$ we find from relations (14.82*) and (14.83*) that for arbitrarily small $\rho > 0$ there hold inequalities $f(M') > f(M_0)$ and $f(M'') < f(M_0)$ which prove that there is no extremum at point M_0.

The proof of Theorem 14.16 is complete.

Remark 1. If the second differential of the function $u = f(M)$ twice differentiable at the given point M_0 of a possible extremum is a semi-definite quadratic form, it cannot be stated for sure whether there is a local extremum at that point or not.

Thus, for instance, the second differential at a point of a possible extremum $M_0(0, 0)$ of each of the two functions $u_1 = x^3 + y^3$ and $u_2 = x^4 + y^4$ is identically equal to zero (that is, it is a semi-definite quadratic form), but only the second of the indicated functions has a local extremum at that point.

To judge upon the existence of a local extremum in the case where the second differential is a semi-definite quadratic form, it is necessary to employ differentials of higher orders, but problems of this kind are not treated within the framework of the present course.

Remark 2. *The requirement* $d^2u\,|_{M_0} \geqslant 0$ ($d^2u\,|_{M_0} \leqslant 0$ *respectively*) *is a necessary condition of a local minimum (maximum) at point* M_0 *of the function* $u = f(M)$ *twice differentiable at that point.*

Indeed, assume for the sake of definiteness that $u = f(M)$ possesses a local minimum at point M_0, but the condition $d^2u\,|_{M_0} \geqslant 0$ is not satisfied. Then there are $h_1, h_2, \ldots, h_m$ such that

$$d^2u\,|_{M_0}=\sum_{i=1}^{m}\sum_{k=1}^{m}\frac{\partial^2 f}{\partial x_i\,\partial x_k}(M_0)\,h_ih_k<0.$$

Let us consider the function $F(t) = f(\overset{\bullet}{x}_1 + th_1, \overset{\bullet}{x}_2 + th_2, \ldots, \overset{\bullet}{x}_m + th_m)$ which is obviously specified for all t sufficiently

small in absolute value. The function $F(t)$ must have a local minimum at point $t = 0$, which comes into a contradiction with the condition

$$F''(0) = d^2u\,|_{M_0} < 0.$$

14.6.3. The case of a function of two variables. The problem that is often encountered in practical applications is that of an extremum of a function of two variables $u = f(x, y)$. We shall now give the results relating to this case.

Let us denote the partial derivatives $\frac{\partial^2 u}{\partial x^2}$, $\frac{\partial^2 u}{\partial x\,\partial y}$, $\frac{\partial^2 u}{\partial y^2}$ at some point $M_0(\overset{\circ}{x}, \overset{\circ}{y})$ as a_{11}, a_{12}, a_{22} respectively. Then the following **assertion** holds true.

Assume that the function $u = f(x, y)$ of two variables is once differentiable in the neighbourhood of point $M_0(\overset{\circ}{x}, \overset{\circ}{y})$ and twice differentiable at point M_0 itself, and let point M_0 be a point of a possible extremum. Then, if the condition $a_{11} \cdot a_{22} - a_{12}^2 > 0$ is satisfied at point M_0, the function $u = f(x, y)$ possesses a local extremum at that point (maximum at $a_{11} < 0$ and minimum at $a_{11} > 0$). Now if $a_{11} \cdot a_{22} - a_{12}^2 < 0$ at point M_0, the function $u = f(x, y)$ does not have a local extremum at that point.*

Proof. The validity of the first part of the assertion we have formulated follows directly from Theorem 14.16 and Sylvester's theorem on the definiteness of sign of a quadratic form, since

$$A_1 = a_{11}, \quad A_2 \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11} \cdot a_{22} - a_{12}^2.$$

We shall prove the second part of the assertion. Thus, let the inequality $a_{11} \cdot a_{22} - a_{12}^2 < 0$ be valid at point M_0. We shall prove that in this case the second differential d^2u constitutes at point M_0 an **alternating** form. Let us first discuss the case $a_{11} \neq 0$. Using the notation

$$\rho = \sqrt{(x - \overset{\circ}{x})^2 + (y - \overset{\circ}{y})^2}, \quad h_1 = \frac{x - \overset{\circ}{x}}{\rho}, \quad h_2 = \frac{y - \overset{\circ}{y}}{\rho},$$

introduced above, we shall obtain the following expression for the second differential:

$$d^2u\,|_{M_0} = \rho^2 \sum_{i=1}^{2} \sum_{k=1}^{2} a_{ik} h_i h_k = \rho^2 (a_{11} h_1^2 + 2a_{12} h_1 h_2 + a_{22} h_2^2) =$$

$$= \frac{\rho^2}{a_{11}} [(a_{11} h_1 + a_{12} h_2)^2 + (a_{11} \cdot a_{22} - a_{12}^2) h_2^2].$$

* The case $a_{11} \cdot a_{22} - a_{12}^2 = 0$ needs a further investigation.

It is easy to verify that at $h_1=1$, $h_2=0$ and at $h_1=\frac{a_{12}}{\sqrt{a_{11}^2+a_{12}^2}}$, $h_2=\frac{a_{11}}{\sqrt{a_{11}^2+a_{12}^2}}$ *, the differential $d^2u|_{M_0}$ has unlike signs, that is, it is an alternating form and, therefore, in accordance with Theorem 14.16, the function does not possess a local extremum at point M_0.

Now let us consider the case $a_{11}=0$. Then it follows from the condition $a_{11}\cdot a_{22}-a_{12}^2<0$ that $a_{12}\neq 0$. Consequently, the expression for $d^2u|_{M_0}$ written above yields

$$d^2u\,|_{M_0}=\rho^2h_2\,(2a_{12}h_1+a_{22}h_2). \tag{14.84}$$

Suppose $h_1\neq 0$ and the quantity h_2 is so small (it follows from the condition $h_1^2+h_2^2=1$ that such a choice of h_1 and h_2 is possible), that the expression $(2a_{12}h_1+a_{22}h_2)$ retains the sign of the quantity $2a_{12}h_1$. Then it follows from formula (14.84) that $d^2u\,|_{M_0}$ has different signs at $h_2>0$ and $h_2<0$, i. e. the function $u=f(x, y)$ does not possess a local extremum at point M_0. The proof of the assertion is complete.

14.6.4. Examples of investigating functions for an extremum.

(1) Find the points of a local extremum of a function of m variables

$$u=\lambda x_1^2+x_2^2+\ldots+x_m^2+2x_2+\ldots+2x_m, \tag{14.85}$$

where λ is a nonzero real number.

To seek for the points of a possible extremum, we obtain the following equations:

$$\frac{\partial u}{\partial x_1}=2\lambda x_1=0,\quad \frac{\partial u}{\partial x_2}=2x_2+2=0,\ \ldots,\ \frac{\partial u}{\partial x_m}=2x_m+2=0. \tag{14.86}$$

We infer from equations (14.86) that the only point of a possible extremum is the point $M_0\,(0, -1, \ldots, -1)$.

To investigate the function (14.85) at that point M_0 with the aid of the sufficient conditions of an extremum, we shall calculate the second differential

$$d^2u\,|_{M_0}=2\lambda\,(dx_1)^2+2\,(dx_2)^2+\ldots+2\,(dx_m)^2. \tag{14.87}$$

It is evident, that at $\lambda>0$ all the values of the second differential (14.87), at $dx_1, dx_2, \ldots, dx_m$ which are not simultaneously zero, are strictly positive, i.e. at $\lambda>0$ the second differential (14.87) is a positive definite quadratic form. Therefore, at $\lambda>0$ the function (14.85) has a local minimum at point $M_0\,(0, -1, \ldots, -1)$.

At $\lambda<0$ the second differential (14.87) is positive at $dx_1=0, \ldots$ $\ldots, dx_{m-1}=0, dx_m=1$ and negative at $dx_1=1, dx_2=0, \ldots$

* In this case ρ can be arbitrarily small. The condition $h_1^2+h_2^2=1$ is satisfied.

$\ldots, dx_m = 0$. This signifies that at $\lambda < 0$ the second differential (14.87) is an alternating quadratic form. Therefore, at $\lambda < 0$ the function (14.85) does not have a local extremum at point $M_0(0, -1, \ldots, -1)$.

(2) Given on the plane are n points $M_k(a_k, b_k)$, $k=1, 2, \ldots, n$, at which the masses $m_k > 0$ are concentrated. It is required to find a point $M_0(x_0, y_0)$ on that plane, relative to which the moment of inertia of the indicated system of particles is minimum.

Since the moment of inertia of the indicated system of particles with respect to point $M(x, y)$ is equal to

$$I(x, y) = \sum_{k=1}^{n} m_k [(x-a_k)^2 + (y-b_k)^2], \tag{14.88}$$

the problem reduces to seeking point $M_0(x_0, y_0)$ at which the function (14.88) attains its minimum value.

To find the points of a possible extremum of the function (14.88), we obtain the following equations:

$$\frac{\partial I}{\partial x} = 2\sum_{k=1}^{n} m_k (x-a_k) = 0, \quad \frac{\partial I}{\partial y} = 2\sum_{k=1}^{n} m_k (y-b_k) = 0. \tag{14.89}$$

We infer from equations (14.89) that the only point of a possible extremum of the function (14.88) is point $M_0(x_0, y_0)$, whose coordinates are

$$x_0 = \frac{m_1a_1 + m_2a_2 + \ldots + m_na_n}{m_1 + m_2 + \ldots + m_n}, \quad y_0 = \frac{m_1b_1 + m_2b_2 + \ldots + m_nb_n}{m_1 + m_2 + \ldots + m_n}. \tag{14.90}$$

Since $a_{11} = \frac{\partial^2 I}{\partial x^2} = 2\sum_{k=1}^{n} m_k > 0$, $a_{12} = \frac{\partial^2 I}{\partial x \partial y} = 0$, $a_{22} = \frac{\partial^2 I}{\partial y^2} = 2\sum_{k=1}^{n} m_k > 0$, it follows that $a_{11} \cdot a_{22} - a_{12}^2 > 0$, and according to the statement proved in 4.6.3 the function (14.88) possesses a local minimum at point $M_0(x_0, y_0)$ with coordinates (14.90). It is easy to ascertain that the value $I(x, y)$ at that point is minimum. We wish to point out in conclusion that formulas (14.90) determine the coordinates of the centre of gravity of the system of particles in question.

14.7. THE GRADIENT METHOD OF SEEKING AN EXTREMUM OF A STRONGLY CONVEX FUNCTION

We present here the theory of the gradient method, widely used in practice, for seeking an extremum of a strongly convex function.

The method is based on a very simple idea. For approximate

determination of a point of minimum of a function of m variables, use is made of the fact that the gradient of this function has a direction coinciding with the direction of the greatest increase of that function. Hence, at each point $x_0 = (\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$ the vector $-\operatorname{grad} f(x_0)$ is directed towards the greatest decrease of the function $f(x) = f(x_1, x_2, \ldots, x_m)$, and this is the reason to suppose that if, proceeding from some zero approximation $x_0 = (\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$, we construct the kth approximation $x_k = (\overset{k}{x}_1, \overset{k}{x}_2, \ldots, \overset{k}{x}_m)$ by the recurrence formula

$$x_{k+1} = x_k - \alpha \operatorname{grad} f(x_k),$$

then for a sufficiently small positive α the sequence of points $\{x_k\}$ will converge to the point of minimum of the function $f(x)$.

The present subsection is devoted to the strict realization of this simple idea.

14.7.1. Convex sets and convex functions. Assume that $x_1 = (\overset{1}{x}_1, \overset{1}{x}_2, \ldots, \overset{1}{x}_m)$ and $x_2 (\overset{2}{x}_1, \overset{2}{x}_2, \ldots, \overset{2}{x}_m)$ are two points of an m-dimensional Euclidean space E^m which may be treated as the vectors in E^m with the corresponding coordinates.

We shall term as the *segment* connecting points x_1 and x_2 a set of points of the space E^m of the form $x_1 + t(x_2 - x_1)$, where t is any number from the closed interval $0 \leqslant t \leqslant 1$.

We shall denote the segment connecting points x_1 and x_2 as $\overline{x_1 x_2}$.

Definition 1. *The set Q of points of the space E^m is said to be* **convex** *if it possesses the following property: whatever the two points x_1 and x_2 belonging to the set Q, the segment $\overline{x_1 x_2}$ connecting them also belongs to that set.*

Cited as an example of a convex set in the space E^m can be an m-dimensional sphere (no matter if open or closed) or a half-space $x_m \geqslant 0$ (i.e. the set of all points $(x_1, x_2, \ldots, x_m)$ of the space E^m, whose m-th coordinate satisfies the condition $x_m \geqslant 0$).

An example of the set Q which is not convex is a complement to an m-dimensional sphere or an m-dimensional sphere with at least one point deleted.

Assume that Q is some set of points of the space E^m and x is any fixed point of E^m.

We shall term as the *distance* from point x to the set Q the greatest lower bound of the distances from x to various points of Q.

We shall use the symbol $\rho(x, Q)$ to designate the distance from point x to the set Q.

Thus, in accordance with the definition

$$\rho(x, Q) = \inf_{y \in Q} \rho(x, y).$$

For any set Q of the space E^m and for any point x of E^m there is a distance $\rho(x, Q)$*. In particular, if point x belongs to the set Q, then $\rho(x, Q) = 0$.

However, there is not always a point y of the set Q such that $\rho(x, y) = \rho(x, Q)$.

Thus, for instance, if the set Q is an *open* m-dimensional sphere and x is a point of E^m lying outside the sphere, then there is no point y of the set Q such that $\rho(x, y) = \rho(x, Q)$ (since the inequality $\rho(x, y) > \rho(x, Q)$ is valid for all points y of the open sphere Q).

If, nonetheless, there is a point y of the set Q such that $\rho(x, y) = \rho(x, Q)$, then y is called the *projection of point x on the set Q*.

The projection of point x on the set Q will be denoted by $P_Q(x)$.

It should be emphasized that if point x belongs to the set Q, then $P_Q(x) = x$.

The projection $P_Q(x)$ of point x on the set Q is specified by the relation

$$\rho(x, P_Q(x)) = \rho(x, Q) = \inf_{y \in Q} \rho(x, y).$$

It pays to notice that there may be several projections of point x on the set Q. Thus, for instance, if Q is an m-dimensional sphere with centre at point x, then any point Q is a projection of point x on the set Q.

The following lemma, however, holds true.

Lemma 1. *If the set Q of the space E^m is convex and closed and x is any point of E^m, then there is a projection of point x on the set Q and that projection is unique.*

Proof. Let us first prove the existence of *at least one* projection of point x on the set Q. We shall denote the distance from point x to the set Q by $\rho(x, Q)$. By the definition of $\rho(x, Q)$ as the greatest lower bound $\inf_{y \in Q} \rho(x, y)$, there is a sequence $\{y_n\}$ of points of the set Q such that $\rho(x, y_n) \to \rho(x, Q)$.

By the definition of the limit of a number sequence, for any $\varepsilon > 0$ all the terms y_n, beginning with some number onwards, satisfy the relation

$$\rho(x, Q) - \varepsilon < \rho(x, y_n) < \rho(x, Q) + \varepsilon.$$

Hence it follows that at any case the sequence $\{y_n\}$ of points of the space E^m is bounded and, therefore, by virtue of Bolzano-Weierstrass's theorem (see 14.2.2) there can be isolated from this sequence a convergent subsequence $\{y_{k_n}\}$ where $n = 1, 2, \ldots$. Let us denote the limit of the subsequence $\{y_{k_n}\}$ by y. Since the set Q

* Since the set $\rho(x, y)$ for various y, belonging to Q, is always bounded below (say, by the number zero).

is closed, point y belongs to that set. It remains to prove that

$$\rho(x, y) = \rho(x, Q) = \lim_{n\to\infty} \rho(x, y_{kn}).$$

To prove this, let us note that by virtue of the inequalities of a triangle $\rho(x, y_{kn}) \leqslant \rho(x, y) + \rho(y, y_{kn})$ and $\rho(x, y) \leqslant \rho(x, y_{kn}) + \rho(y_{kn}, y)$ the following inequality holds true:

$$|\rho(x, y_{kn}) - \rho(x, y)| \leqslant \rho(y, y_{kn}).$$

This relation and the convergence of the subsequence $\{y_{kn}\}$ to y yield $\lim_{n\to\infty} \rho(x, y_{kn}) = \rho(x, y)$, i.e. $\rho(x, Q) = \rho(x, y)$.

Thus we have completed the proof of the existence of at least one projection of point x on the set Q.

We shall now prove that there is *only one* projection of point x on the set Q. Suppose that there are two *different* projections y_1 and y_2 of point x on the set Q. Since the set Q is convex, the entire segment $\overline{y_1y_2}$, joining points y_1 and y_2, belongs to the set Q. In particular, the set Q contains the middle point $\frac{y_1+y_2}{2}$ of the indicated segment. We shall make sure that the *distance* $\rho\left(x, \frac{y_1+y_2}{2}\right)$ *from point* x *to the middle point of the segment* $\overline{y_1y_2}$ *is strictly less than the distance* $\rho(x, y_1) = \rho(x, y_2)$.

We shall not consider the trivial case of $\frac{y_1+y_2}{2} = x$. In this case $\rho\left(x, \frac{y_1+y_2}{2}\right) = 0$ whereas $\rho(x, y_1) = \rho(x, y_2) > 0$, since if otherwise (i.e. if $\rho(x, y_1) = \rho(x, y_2) = 0$), both points y_1 and y_2 would coincide with x and could not be different. Thus we see that in the trivial case $\frac{y_1+y_2}{2} = x$ the inequality

$$\rho\left(x, \frac{y_1+y_2}{2}\right) < \rho(x, y_1) = \rho(x, y_2) \tag{14.91}$$

is obvious.

Let us now prove inequality (14.91) for the case where $\frac{y_1+y_2}{2} \neq x$.

Using the properties of the scalar product of two vectors of the space E^m, we shall obtain the relation

$$\rho^2\left(x, \frac{y_1+y_2}{2}\right) = \left(\frac{y_1+y_2}{2} - x, \frac{y_1+y_2}{2} - x\right) =$$

$$= \left(\frac{y_1-x}{2} + \frac{y_2-x}{2}, \frac{y_1-x}{2} + \frac{y_2-x}{2}\right) =$$

$$= \frac{1}{4}[(y_1-x, y_1-x) + 2(y_1-x, y_2-x) + (y_2-x, y_2-x)]. \tag{14.92}$$

Let us verify the validity of the strict inequality

$$|(y_1-x,\ y_2-x)|<\sqrt{(y_1-x,\ y_1-x)}\cdot\sqrt{(y_2-x,\ y_2-x)}. \quad (14.93)$$

To do that, let us make use of the fact that for any two vectors $\mathbf{a}$ and $\mathbf{b}$ of the space E^m that are not collinear (that is, such that $\mathbf{a}\neq\lambda\mathbf{b}$ for any real λ), the strict inequality of Cauchy-Bunyakovsky*

$$|(\mathbf{a},\ \mathbf{b})|\leqslant\sqrt{(\mathbf{a},\ \mathbf{a})\cdot(\mathbf{b},\ \mathbf{b})}$$

holds true. This means that to prove inequality (14.93), it is sufficient to make sure that the vectors y_1-x and y_2-x are not collinear, i.e. ascertain the fact that the equality

$$y_1-x=\lambda(y_2-x) \quad (14.94)$$

cannot hold true for any real λ. Were equality (14.94) valid for a λ such that $|\lambda|\neq 1$, the equality $\rho(y_1,\ x)=\rho(y_2,\ x)$ would be impossible. The validity of equality (14.94) for $\lambda=+1$ would contradict the fact that points y_1 and y_2 are distinct. And, finally, the validity of equality (14.94) for $\lambda=-1$ would mean that $\frac{y_1+y_2}{2}=x$, but we have excluded that case from consideration.

Thus we see that equality (14.94) is not valid for any real λ and that is the proof of inequality (14.93).

Comparing equality (14.92) and inequality (14.93), we obtain

$$\rho^2\left(x,\frac{y_1-y_2}{2}\right)<\frac{1}{4}[(y_1-x,\ y_1-x)+$$
$$+2\sqrt{(y_1-x,\ y_2-x)\,(y_2-x,\ y_2-x)}+(y_2-x,\ y_2-x)]=$$
$$=\frac{1}{4}[\sqrt{(y_1-x,\ y_1-x)}+\sqrt{(y_2-x,\ y_2-x)}]^2=$$
$$=\frac{1}{4}[\rho(x,\ y_1)+\rho(x,\ y_2)]^2=\rho^2(x,\ y_1)=\rho^2(x,\ y_2).$$

Thus we have completed the proof of inequality (14.91). But this inequality signifies that there is a point $\frac{y_1+y_2}{2}$ of the set Q which is closer to x than points y_1 and y_2, and this contradicts the fact that each of the points y_1 and y_2 is the projection of point x on the set Q, i.e. is the greatest lower bound of the distance $\rho(x,\ y)$ for various y belonging to Q.

The contradiction we have obtained shows that the assumption as to the existence of two different projections y_1 and y_2 of point x on the set Q is erroneous.

---

* Indeed, for $\mathbf{a}\neq\lambda\mathbf{b}$ the vector $(\mathbf{a}-\lambda\mathbf{b})$ is nonzero. Therefore, the quadratic trinomial $(\mathbf{a}-\lambda\mathbf{b},\ \mathbf{a}-\lambda\mathbf{b})=(\mathbf{a},\ \mathbf{a})-2\lambda(\mathbf{a},\ \mathbf{b})+\lambda^2(\mathbf{b},\ \mathbf{b})$ is strictly positive and its discriminant $4[(\mathbf{a},\ \mathbf{b})^2-(\mathbf{a},\ \mathbf{a})\cdot(\mathbf{b},\ \mathbf{b})]$ is strictly negative.

The proof of Lemma 1 is complete.

Let us now define a convex function.

Definition 2. *The function $f(x)$ defined on the convex set Q of the space E^m is said to be* **convex downwards** *or simply* **convex** *on that set if for any two points x_1 and x_2 of the set Q and for any real number t from the closed interval $0 \leqslant t \leqslant 1$ there holds an inequality*

$$f[x_2 + t(x_2 - x_1)] \leqslant f(x_1) + t[f(x_2) - f(x_1)]. \tag{14.95}$$

Definition 3. *The function $f(x)$ defined on the convex set Q of the space E^m is said to be* **strictly convex** *on that set if for any two points x_1 and x_2 of the set Q and for any real number t from the interval $0 < t < 1$ there holds a strict inequality*

$$f[x_2 + t(x_2 - x_1)] < f(x_1) + t[f(x_2) - f(x_1)]. \tag{14.96}$$

It is clear that every function $f(x)$ strictly convex on the set Q is convex on that set.

It is easy to establish the sufficient condition of convexity (strict convexity respectively) of the function $f(x)$ twice differentiable on the convex set Q.

In what follows we shall everywhere assume that the set Q possesses at least one interior point.

Lemma 2. *Assume that the function $f(x)$ is defined and twice differentiable on the convex set Q. Then, for that function to be convex (strictly convex) on the set Q, it is sufficient that the second differential d^2f of that function should be a positive semidefinite (strictly positive definite) quadratic form at all the points of Q.*

Proof. Suppose x_1 and x_2 are any two fixed points of the set Q. Let us consider, on the closed interval $0 \leqslant t \leqslant 1$, the following function of one independent variable t:

$$F(t) = f[x_1 + t(x_2 - x_1)] - f(x_1) - t[f(x_2) - f(x_1)]. \tag{14.97}$$

Recall that the second differential d^2f of the function $f(x) = f(x_1, x_2, \ldots, x_m)$ of m independent variables $x_1, x_2, \ldots, x_m$ at the given point $x = (x_1, x_2, \ldots, x_m)$ is equal to*

$$d^2f(x) = \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{\partial^2 f}{\partial x_i \partial x_k}(x) \cdot \Delta x_i \Delta x_k. \tag{14.98}$$

Differentiating the function (14.97) two times with respect to t in accordance with the rule of differentiating a composite function, we obtain

$$F''(t) = \sum_{i=1}^{m} \sum_{k=1}^{m} \frac{\partial^2 f}{\partial x_i \partial x_k}[x_1 + t(x_2 - x_1)]\,(\overset{2}{x_i} - \overset{1}{x_i})\,(\overset{2}{x_k} - \overset{1}{x_k}), \tag{14.99}$$

* See 14.5.2.

where $(\overset{1}{x}_1, \overset{1}{x}_2, \ldots, \overset{1}{x}_m)$ and $(\overset{2}{x}_1, \overset{2}{x}_2, \ldots, \overset{2}{x}_m)$ are the coordinates of points x_1 and x_2 respectively.

Comparing relations (14.98) and (14.99), we shall ascertain the validity of the equality

$$F''(t) = d^2f[x_1 + t(x_2 - x_1)], \tag{14.100}$$

where in the expression for d^2f the increments Δx_i are taken equal to $\overset{2}{x}_i - \overset{1}{x}_i$.

For the sake of definiteness, the arguments that will follow will concern the case where the second differential d^2f at all the points of Q is a positive semidefinite quadratic form. Then, for all t from the closed interval $0 \leqslant t \leqslant 1$ the right-hand (and, consequently, the left-hand) side of (14.100) is nonnegative, that is, for all t from the closed interval $0 \leqslant t \leqslant 1$

$$F''(t) \geqslant 0. \tag{14.101}$$

By virtue of Definition 2 and relation (14.97), it suffices to prove that for all t from the closed interval $0 \leqslant t \leqslant 1$ there holds an inequality

$$F(t) \leqslant 0. \tag{14.102}$$

To prove inequality (14.102), let us make use of inequality (14.101) and the easily verifiable equalities

$$F(0) = 0, \quad F(1) = 0. \tag{14.103}$$

Let us suppose that there exists, within the closed interval $0 \leqslant t \leqslant 1$, at least one point t in which $F(t) > 0$. Then the function $F(t)$ attains its maximum value on the closed interval $0 \leqslant t \leqslant 1$ at some interior point t_0 of that interval, with $F(t_0) > 0$. At the point t_0 the function $F(t)$ possesses a local maximum and, therefore, $F'(t_0) = 0$. But it follows from inequality (14.101) that the derivative $F'(t)$ does not decrease throughout the closed interval $0 \leqslant t \leqslant 1$ and, consequently, on the closed interval $t_0 \leqslant t \leqslant 1$. From this and from the condition $F'(t_0) = 0$ it follows that the derivative $F'(t)$ is nonnegative everywhere on the closed interval $t_0 \leqslant t \leqslant 1$ and, therefore, the function $F(t)$ does not decrease on that interval. This leads to the inequality

$$F(1) \geqslant F(t_0) > 0,$$

which contradicts the second relation (14.103).

The contradiction we have obtained proves the assumption concerning the existence of at least one point t, within the closed interval $0 \leqslant t \leqslant 1$, at which $F(t) > 0$, to be erroneous, that is, proves the validity of inequality (14.102) everywhere on the closed interval $0 \leqslant t \leqslant 1$.

And this completes the proof of the first part of the lemma (on the convexity of $f(x)$ under the condition that d^2f is a positive semidefinite quadratic form).

By a complete analogy we can prove the second part of the lemma (on the strict convexity of $f(x)$ under the condition that d^2f is a strictly positive definite quadratic form). Proceeding from inequality (14.101) which is valid this time with the $>$ sign and from equalities (14.103) and assuming that there is at least one point t, within the closed interval $0 \leqslant t \leqslant 1$, at which $F(t) \geqslant 0$, we shall come to the conclusion that $F(t)$ possesses a point of a local maximum t_0 within the closed interval $0 \leqslant t \leqslant 1$, with $F(t_0) \geqslant 0$. But then, since $F'(t_0) = 0$ we shall find from (14.101) that $F'(t) > 0$ everywhere on the half interval $t_0 < t \leqslant 1$, and this signifies that $F(1) > F(t_0) \geqslant 0$.

This is again at variance with the second relation (14.103), which proves that $F(t) < 0$ everywhere on the interval $0 < t < 1$, that is, proves the strict convexity of $f(x)$ on the set Q.

This completes the proof of Lemma 2.

The lemma we have proved naturally leads to the consideration of the following, still narrower, class of functions convex on the convex set Q and twice differentiable on that set.

Definition 4. *The function $f(x)$ twice differentiable on the convex set Q is said to be* **strongly convex** *on that set if there exist two positive constants k_1 and k_2 such that the second differential d^2f of this function, specified by relation* (14.98), *at all points x of the set Q satisfies the inequalities*

$$k_1(\Delta x)^2 \leqslant d^2f \leqslant k_2(\Delta x)^2. \tag{14.104}$$

(In these inequalities Δx denotes the vector with the coordinates $(\Delta x_1, \Delta x_2, \ldots, \Delta x_m)$ and the symbol $(\Delta x)^2$ is used for the scalar square of this vector.)

It immediately follows from the left-hand inequality (14.104) that the second differential of a strongly convex function is a strictly positive definite function at all the points of the set Q and, therefore (by virtue of Lemma 2), *a function strongly convex on the set Q is obviously strictly convex on that set.*

Since this class of strongly convex functions is sufficiently wide and very important for applications, we shall restrict the presentation of the theory of the gradient method of seeking a minimum to this class of functions.

We shall begin with elucidating the problem of the existence and the uniqueness of a minimum.

14.7.2. The existence of a minimum of a strongly convex function and the uniqueness of the minimum of a strictly convex function. Suppose the function $f(x)$ is defined on the convex set Q. We shall say that the function possesses a *local minimum* at point x_0 of the

set Q if there exists a δ-neighbourhood of point x_0 such that the value $f(x_0)$ is the smallest among the values $f(x)$ of this function at all the points of intersection of the δ-neighbourhood of x_0 and the set Q.

According to this definition, the concept of a local minimum includes the points of an end-point minimum of the function $f(x)$ on the boundary of the set Q.

Thus, in accordance with the definition we have given, we can divide the points of minimum into the points of an interior local minimum (for the case where these points are interior points of Q) and the points of an end-point local minimum (for the case where these points are boundary points of Q).

To study the problem of the existence and the uniqueness of the point of a local minimum, we shall need the following auxiliary theorem.

Lemma 3. *Assume that a differentiable convex function $f(x)$ is given on the convex set Q. For this function to have a local minimum at point x_0 of the set Q, it is necessary and sufficient that for any vector Δx for which the point $x_0 + \Delta x$ belongs to the set Q, the following inequality should hold true**:

$$(\operatorname{grad} f(x_0),\ \Delta x) \geqslant 0. \tag{14.105}$$

Proof. (1) **Necessity.** By virtue of the assertion proved in 14.4.6 the left-hand side of (14.105) is equal to the product of the derivative of the function $f(x)$ at point x_0 along the direction of the vector Δx by the length $|\Delta x|$ of this vector:

$$(\operatorname{grad} f(x_0),\ \Delta x) = \frac{\partial f}{\partial e}(x_0)\,|\Delta x|, \tag{14.106}$$

where $e = \dfrac{\Delta x}{|\Delta x|}$ is a unit vector in the direction of Δx.

Since x_0 is a point of a local minimum of the function $f(x)$, the derivative $\dfrac{\partial f}{\partial e}(x_0)$ in any direction of $e = \dfrac{\Delta x}{|\Delta x|}$ is nonnegative (to be more precise, it is equal to zero when x_0 is a point of the interior local extremum and nonnegative when x_0 is a point of an end-point local extremum).

Thus we have found that the right-hand side of (14.106) (and, consequently, the left-hand side of (14.105)) is nonnegative, and proved the necessity.

(2) **Sufficiency.** Let inequality (14.105) be valid for any vector Δx for which the point $x_0 + \Delta x$ belongs to Q. We shall prove that point x_0 is a point of a local minimum of the function $f(x)$.

---

* In inequality (14.105) we take the scalar product of the vectors grad $f(x_0)$ and Δx. For the definition of grad $f(x_0)$ see 14.4.6.

Since by the hypothesis the function $f(x)$ is convex on the set Q, inequality (14.95) is valid for any two points x_1 and x_2 of that set and for any number t belonging to the closed interval $0 \leqslant t \leqslant 1$. Putting $x_1 = x_0$, $x_2 = x_0 + \Delta x$ in this inequality, we can rewrite the inequality in the form

$$f(x_0 + \Delta x) - f(x_0) \geqslant \frac{f(x_0 + t\Delta x) - f(x_0)}{t}. \qquad (14.107)$$

Assuming x_0 and Δx to be fixed, we shall pass to the limit in inequality (14.107) at $t \to 0 + 0$. By the definition of the directional derivative (see 14.4.6), the limit on the right-hand side of (14.107) at $t \to 0 + 0$ is precisely equal to the product appearing on the right-hand side of (14.106). Therefore, by virtue of relations (14.105) and (14.106) this limit is nonnegative. Taking into account that the left-hand side of (14.107) does not depend on t, we receive in the limit, at $t \to 0 + 0$, from inequality (14.107),

$$f(x_0 + \Delta x) - f(x_0) \geqslant 0.$$

The last inequality valid for any vector Δx for which the point $x_0 + \Delta x$ belongs to Q proves that the function $f(x)$ possesses a local minimum at point x_0, and that is the proof of the sufficiency.

Lemma 3 has been completely proved.

Remark 1. It is evident from the proof we have carried out that for the case where point x_0 is an *interior* point of the set Q, i.e. when we mean an interior local minimum, the $\geqslant$ sign in inequality (14.105) in the formulation of Lemma 3 can be replaced by the equality sign.

Remark 2. In proving the necessity in Lemma 3 we have not used the requirement of the convexity of the function $f(x)$. Therefore, the proof of the necessity does not include the requirement of the convexity of the function $f(x)$. In other words, the following *assertion* holds true: *if the function $f(x)$ is differentiable on the convex set Q and possesses a local minimum in the interior (boundary) point x_0 of that set, then for any vector Δx for which point $x_0 + \Delta x$ belongs to Q, there holds an inequality*

$$(\operatorname{grad} f(x_0), \Delta x) = 0 \quad [(\operatorname{grad} f(x_0), \Delta x) \geqslant 0].$$

Let us now consider the question of the uniqueness and the existence of the point of a local minimum.

Theorem (on the uniqueness of a local minimum of a strictly convex function). *If the function $f(x)$ is differentiable and strictly convex on the convex set Q, it can have a local minimum only at one point of that set.*

Proof. Let us suppose that the function $f(x)$ possesses a local minimum at two *different* points x_1 and x_2 of the set Q. Then the condition of convexity (14.95) for the points x_1 and x_2 can be written in

the form

$$f(x_2) - f(x_1) \geqslant \frac{f[x_1 + t(x_2 - x_1)] - f(x_1)}{t} \tag{14.108}$$

(here t is any number from the closed interval $0 \leqslant t \leqslant 1$).

Making points x_1 and x_2 exchange their roles in relation (14.108), we obtain

$$f(x_1) - f(x_2) \geqslant \frac{f[x_2 + t(x_1 - x_2)] - f(x_2)}{t}. \tag{14.109}$$

In the limit, at $t \to 0 + 0$, the right-hand side of (14.108) (the right-hand side of (14.109) respectively) yields a derivative of the function $f(x)$ along the direction of the vector $x_2 - x_1$ (the vector $x_1 - x_2$ respectively) taken at point x_1 (at point x_2 respectively) multiplied by $|x_2 - x_1|$. Both points x_1 and x_2 being the points of a local minimum, the two directional derivatives are nonnegative, that is, the limits of the right-hand sides of (14.108) and (14.109) are both nonnegative at $t \to 0 + 0$.

Thus inequalities (14.108) and (14.109) yield in the limit, at $t \to 0 + 0$,

$$f(x_2) - f(x_1) \geqslant 0,$$
$$f(x_1) - f(x_2) \geqslant 0.$$

Comparing the last two inequalities, we find that $f(x_1) = f(x_2)$. Making use of the equality $f(x_1) = f(x_2)$, we find from the condition of the strict convexity (14.96) that

$$f[x_1 + t(x_2 - x_1)] < f(x_1) \tag{14.110}$$

for *all* t belonging to the interval $0 < t < 1$.

Inequality (14.110) contradicts the fact that the function $f(x)$ has a local minimum at point x_1 (at point $x_1 + t(x_2 - x_1)$ arbitrarily close to point x_1 for small t, the function $f(x)$ has the value smaller than the value $f(x_1)$).

The contradiction we have obtained shows that our supposition that the function $f(x)$ has a local minimum at two different points of the set Q is erroneous, and this is what we wished to prove.

We shall now prove the existence of a local minimum under stronger restrictions than the uniqueness.

Theorem (on the existence of a local minimum of a strongly convex function. *If the function $f(x)$ is strongly convex on the closed convex set Q, then this function has a point x_0 of a local minimum on the set Q*.*

* Since the function $f(x)$ strongly convex on the convex set Q is strictly convex on that set, in accordance with the previous theorem point x_0 is the *only* point of a local minimum.

Proof. First we must point out that the theorem is obviously valid for the case when the set Q is not only closed and convex but also *bounded*. In this case, in accordance with Weierstrass's second theorem (see Theorem 14.7), the function $f(x)$, being at any rate continuous on the set Q, attains its minimum value on Q at some point x_0 of that set. This point x_0 is precisely the point of a local minimum.

It remains to prove the theorem for the case where the closed convex set Q *is not bounded*. In this case we fix some interior point x_1 of the set Q and expand the function $f(x)$ by Taylor's formula with centre at point x_1 taking the remainder $R_2(x)$ in Lagrange's form* (see 14.5.3). The expansion will have the form

$$f(x) = f(x_1) + df(x_1) + \frac{1}{2} d^2 f[x_1 + \theta(x - x_1)], \tag{14.111}$$

where θ is a number belonging to the interval $0 < \theta < 1$, so that point $x_1 + \theta(x - x_1)$ belongs to the segment connecting points x_1 and x**.

If we denote the vector $x - x_1$ by Δx, then the following equality will be valid for $df(x_1)$:

$$df(x_1) = (\operatorname{grad} f(x_1), \Delta x).$$

This equality yields

$$|df(x_1)| \leqslant |\operatorname{grad} f(x_1)| \cdot |\Delta x|. \tag{14.112}$$

Next, making use of the left-hand inequality in the definition of strong convexity (14.104), we shall arrive at

$$d^2 f[x_1 + \theta(x - x_1)] \geqslant k_1 (\Delta x)^2. \tag{14.113}$$

We infer from relations (14.111)-(14.113) that

$$f(x) - f(x_1) \geqslant -|df(x_1)| + \frac{1}{2} d^2 f[x_1 + \theta(x - x_1)] \geqslant$$

$$\geqslant -|\operatorname{grad} f(x_1)| \, |\Delta x| + \frac{k_1}{2} |\Delta x|^2,$$

so that

$$f(x) - f(x_1) \geqslant |\Delta x| \left[\frac{k_1}{2} |\Delta x| - |\operatorname{grad} f(x_1)| \right] \tag{14.114}$$

Recalling that point x_1 is fixed and the quantity $|\operatorname{grad} f(x_1)|$ is some fixed number, we can obviously choose a positive number R

---

* We take into account that the function $f(x)$ strongly convex on the set Q is twice differentiable on that set.

** Whatever the point x of the set Q, the segment connecting points x_1 and x_2 belongs to the set Q by virtue of the convexity of that set. It was pointed out in the footnote to Taylor's theorem 14.15 that we can take any convex neighbourhood of the centre of expansions as the neighbourhood of that centre, i.e. we can take the entire set Q.

large enough for the expression in brackets in (14.114) to be positive for $|\Delta x| > R$.

This means that for $|\Delta x| > R$ the inequality $f(x) > f(x_1)$ holds true, that is, everywhere outside the closed sphere C_R of radius R with centre at point x_1 the values $f(x)$ exceed the value $f(x_1)$ (at the centre of the sphere).

Let us denote by Q_R the intersection of the set Q with the sphere C_R. Since both sets Q and C_R are convex and closed, their intersection Q_R is convex and closed too. Since, in addition, the set Q_R is bounded, the function $f(x)$, as was proved above, possesses a single point x_0 of a local minimum on the set Q_R.

Inasmuch as we have proved that at all the points of Q lying outside of Q_R the values $f(x)$ exceed the value $f(x_1)$, these values are sure to exceed $f(x_0)$, i.e. point x_0 is a point of a local minimum of $f(x)$ on the entire set Q as well.

The proof of the theorem is complete.

14.7.3. Seeking a minimum of a strongly convex function. We have proved that the strongly convex function $f(x)$ specified on the closed convex set Q possesses on that set a single point x_0 of a local minimum.

Let us try to construct and substantiate an algorithm for seeking this point x_0.

We fix an arbitrary point x_1 of the set Q and an arbitrary number α satisfying the inequalities

$$0 < \alpha < \frac{2}{k_2} \tag{14.115}$$

where k_2 is a constant from inequality (14.104) specifying the strong convexity of the function $f(x)$.

Proceeding from x_1 as the first approximation, we compile an iterative sequence $\{x_k\}$ with the aid of the recurrence relation

$$x_{k+1} = P_Q(x_k - \alpha \operatorname{grad} f(x_k)), \quad k = 1, 2, \tag{14.116}$$

We shall now prove the following assertion.

Fundamental Theorem. *Let the function $f(x)$ be strongly convex on the closed convex set Q and let x_1 be an arbitrary point of the set Q. Then the iterative sequence $\{x_k\}$ specified by the recurrence relation* (14.116) *for any α satisfying inequality* (14.115), *converges to point x_0 of the local minimum of the function $f(x)$.*

It should be emphasized that this theorem provides an algorithm for seeking any (interior or end-point) local minimum of the function $f(x)$ which is strongly convex on the arbitrary (not necessarily bounded) closed convex set Q.

We preface the proof of the fundamental theorem with four lemmas.

Lemma 4. If Q is a closed convex set of points of E^m, x is an arbitrary fixed point of E^m and y is an arbitrary point of Q, then

$$(x - P_Q(x),\ y - P_Q(x)) \leqslant 0. \tag{14.117}$$

Proof. Let us suppose that inequality (14.117) is not valid. Then there is a point y of the set Q such that

$$(x - P_Q(x),\ y - P_Q(x)) > 0. \tag{14.118}$$

It immediately follows from (14.118) that point y does not coincide with $P_Q(x)$.

The set Q being convex, any point $z = P_Q(x) + t(y - P_Q(x))$ of the line segment connecting points $P_Q(x)$ and y belongs to the set Q. Let us calculate the distance between any such point z and point x:

$$\rho^2(z, x) = (x - P_Q(x) - t(y - P_Q(x)),\ x - P_Q(x) - t(y - P_Q(x))) =$$
$$= \rho^2(x, P_Q(x)) - 2t(x - P_Q(x),\ y - P_Q(x)) + t^2\rho^2(y, P_Q(x)). \tag{14.119}$$

Since x and y are fixed and t is any number from the closed interval $0 \leqslant t \leqslant 1$, by virtue of inequality (14.118) we can take t satisfying the inequality

$$0 < t < \frac{2(x - P_Q(x),\ y - P_Q(x))}{\rho^2(y, P_Q(x))}$$

The choice of such a t results in

$$-2t(x - P_Q(x),\ y - P_Q(x)) + t^2\rho^2(y, P_Q(x)) < 0,$$

and we find from (14.119) that

$$\rho^2(z, x) < \rho^2(x, P_Q(x)).$$

The last inequality contradicts the fact that point $P_Q(x)$ is a projection of point x on the set Q: we have found a point z of the set Q which is less distant from x than $P_Q(x)$. The contradiction obtained completes the proof of the lemma.

Lemma 5. Assume that the function $f(x)$ is differentiable and convex on the closed convex set Q. If for some positive α the projection $P_Q(x_0 - \alpha \operatorname{grad} f(x_0))$ of point $x_0 - \alpha \operatorname{grad} f(x_0)$ on the set Q coincides with point x_0 of that set, then the function $f(x)$ has a local minimum at point x_0.

Proof. Making use of Lemma 4, we shall write inequality (14.117) for the points $x = x_0 - \alpha \operatorname{grad} f(x_0)$ and $y = x_0 + \Delta x$, where Δx is any vector for which point $y = x_0 + \Delta x$ belongs to Q. As a result we shall have

$$(x_0 - \alpha \operatorname{grad} f(x_0) - P_Q(x_0 - \alpha \operatorname{grad} f(x_0)),$$
$$x_0 + \Delta x - P_Q(x_0 - \alpha \operatorname{grad} f(x_0))) \leqslant 0.$$

Recalling that $P_Q(x_0 - \alpha \operatorname{grad} f(x_0)) = x_0$, we obtain from the last inequality the following relation:

$$(\operatorname{grad} f(x_0), \Delta x) \geqslant 0.$$

By virtue of Lemma 3, this relation, valid for any vector Δx for which point $x_0 + \Delta x$ belongs to Q, establishes that the function $f(x)$ possesses a local minimum at point x_0. We have proved Lemma 5.

Let us assume that the function $f(x)$ is strongly convex on the bounded closed convex set Q. We shall denote by m the minimum value of $f(x)$ on the set Q and by μ the number, strictly greater than m, so that

$$\mu > m = \min_{x \in Q} f(x).$$

We shall next fix the number ν, strictly greater than μ, and denote by $\hat{Q}$ the subset of points x of the set Q for which

$$\mu \leqslant f(x) \leqslant \nu \tag{14.120}$$

The set $\hat{Q}$ being a subset of the bounded set Q is itself bounded.

Let us verify the fact that the set $\hat{Q}$ is *closed*. Suppose that $\{x_k\}$ is an arbitrary convergent sequence of points of the set $\hat{Q}$. It is required to prove that the limit x_0* of that sequence also belongs to the set $\hat{Q}$. Since every point x_k belongs to the set $\hat{Q}$, the following inequality holds for every number k:

$$\mu \leqslant f(x_k) \leqslant \nu \tag{14.121}$$

In any event, the strictly convex function $f(x)$ is continuous on Q and, therefore, it follows from the convergence of the sequence $\{x_k\}$ to x_0, by the definition of the continuity of a function, that the sequence $\{f(x_k)\}$ converges to the number $f(x_0)$. Since all the terms of the convergent number sequence $\{f(x_k)\}$ satisfy inequalities (14.121), the limit $f(x_0)$ of that sequence also satisfies the inequalities $\mu \leqslant f(x_0) \leqslant \nu$ (see Chap. 3, Corollary 2 of Theorem 3.13), and this precisely means that point x_0 belongs to the set $\hat{Q}$. We have proved that the set $\hat{Q}$ is closed.

Thus we have found that the set $\hat{Q}$ of all points x of the set Q for which inequalities (14.120) hold true, is closed and bounded.

Let us now prove the following lemma.

Lemma 6. *Let the function $f(x)$ be strongly convex on the closed convex set Q, x be any point of Q and α any positive number, the symbol Δx denoting the difference*

$$\Delta x = P_Q(x - \alpha \operatorname{grad} f(x)) - x. \tag{14.122}$$

* Since the initial set Q is closed, the limit x_0 belongs to Q in any case.

Then the following inequality holds true:

$$(\operatorname{grad} f(x),\ \Delta x) \leqslant \frac{1}{\alpha} |\Delta x| \tag{14.123}$$

If, in addition, the set Q is bounded and point x belongs to the subset $\hat{Q}$ of the points of Q for which inequalities (14.120) hold true at $\mu > \min_{x \in Q} f(x)$, then there can be chosen a positive number γ such that the following inequality holds true:

$$|\Delta x| \geqslant \gamma. \tag{14.124}$$

Proof. We shall first prove inequality (14.123). To do that, we shall fix an arbitrary point x of the set Q and, with the aid of Lemma 4, write inequality (14.117), taking point $x - \alpha \operatorname{grad} f(x)$ in it instead of x and point x instead of y. We shall obtain the inequality

$$(x - \alpha \operatorname{grad} f(x) - P_Q(x - \alpha \operatorname{grad} f(x))_{\text{a}}$$
$$x - P_Q(x - \alpha \operatorname{grad} f(x))) \leqslant 0,$$

which, with account taken of the notation $\Delta x = P_Q(x - \alpha \operatorname{grad} f(x)) - x$, can be rewritten as

$$(-\alpha \operatorname{grad} f(x) - \Delta x,\ -\Delta x) \leqslant 0.$$

It follows from the last inequality and from the properties of a scalar product that

$$\alpha (\operatorname{grad} f(x),\ \Delta x) + |\Delta x|^2 \leqslant 0,$$

and this leads to inequality (14.123).

It remains to prove that under the additional assumption to the point that Q is bounded and that x belongs to the subset $\hat{Q}$, there is a $\gamma > 0$ such that inequality (14.124) holds true.

Let us now consider a nonnegative function of point x of the form

$$|\Delta x| = |P_Q(x - \alpha \operatorname{grad} f(x)) - x|. \tag{14.125}$$

We shall make sure that this function is a function of point x continuous on the set Q.

We shall first prove that the vector function $P_Q(x)$ is a continuous function of point x. For our purpose it is sufficient to prove the inequality

$$|P_Q(x + \Delta x) - P(x)| \leqslant |\Delta x|, \tag{14.126}$$

valid for any vectors x and Δx.

By virtue of Lemma 4, the following inequalities hold true:

$$(x - P_Q(x),\ P_Q(x + \Delta x) - P_Q(x)) \leqslant 0,$$
$$(x + \Delta x - P_Q(x + \Delta x),\ P_Q(x) - P_Q(x + \Delta x)) \leqslant 0.$$

Making use of these inequalities and the Cauchy-Bunyakovsky inequality, we shall obtain a chain of relations

$$|P_Q(x+\Delta x)-P_Q(x)|^2=(P_Q(x+\Delta x)-P_Q(x),\ P_Q(x+\Delta x)-$$
$$-P_Q(x))=(P_Q(x+\Delta x)-x,\ P_Q(x+\Delta x)-P_Q(x))+(x-P_Q(x),$$
$$P_Q(x+\Delta x)-P_Q(x))\leqslant(P_Q(x+\Delta x)-x,\ P_Q(x+\Delta x)-P_Q(x))=$$
$$=(P_Q(x+\Delta x)-x-\Delta x,\ P_Q(x+\Delta x)-P_Q(x))+(\Delta x,\ P_Q(x+\Delta x)-$$
$$-P_Q(x))\leqslant(\Delta x,\ P_Q(x+\Delta x)-P_Q(x))\leqslant|\Delta x|\cdot|P_Q(x+\Delta x)-$$
$$-P_Q(x)|,$$

which yields inequality (14.126).

We have thus proved that $P_Q(x)$ is a continuous vector function of point x. It follows from the strong convexity of $f(x)$ on Q that the function α grad $f(x)$ is a vector function of point x also continuous on Q. But then, from the theorem on the continuity of a composite function and the continuity of the difference between continuous functions it follows that the function

$$P_Q(x-\alpha \operatorname{grad} f(x))-x$$

is a vector function of point x continuous on the set Q.

All the more so, the absolute value of the indicated vector function, i.e. the scalar function (14.125), is continuous on the set Q and, consequently, on its subset $\hat{Q}$.

Thus we have found that the function (14.125) is continuous and nonnegative everywhere on the closed bounded set $\hat{Q}$. In that case, in accordance with Weierstrass's second theorem (see Theorem 14.7) this function attains on $\hat{Q}$ its nonnegative minimum value γ which is obviously strictly positive. Were γ equal to zero, there would be a point x_0 on $\hat{Q}$ such that $P_Q(x_0-\alpha \operatorname{grad} f(x_0))-x_0=0$ and that would mean, by virtue of Lemma 5, that at this point x_0 of $\hat{Q}$ the function $f(x)$ possesses the only, on the set Q, local minimum (whereas, by the definition of $\hat{Q}$, this minimum lies outside of $\hat{Q}$). Thus we see that $\gamma>0$ and this is the proof of inequality (14.124).

We have completed the proof of Lemma 6.

Lemma 7. *Assume that the function $f(x)$ is strongly convex on the closed convex set Q, x is any point of Q, α is any number satisfying inequalities* (14.115), *and Δx is a difference of the form of* (14.122). *Then on passing from point x to point $x^*=P_Q(x-\alpha\cdot\operatorname{grad} f(x))$, the value of the function $f(x)$ does not increase, with**

$$f(x)-f(x^*)\geqslant\left(\frac{1}{\alpha}-\frac{k_2}{2}\right)\cdot|\Delta x|^2. \qquad (14.127)$$

* It follows from (14.115) that $\left(\frac{1}{\alpha}-\frac{k_2}{2}\right)>0$.

If, in addition, the set Q is bounded and point x belongs to the subset $\hat{Q}$ of the points of Q for which inequality (14.120) *holds true at* $\mu > \min_{x \in Q} f(x)$, *then inequality* (14.127) *passes into the inequality*

$$f(x) - f(x^*) \geqslant \left(\frac{1}{\alpha} - \frac{k_2}{2}\right) \gamma^2, \tag{14.128}$$

where $\gamma > 0$ *is a constant from Lemma* 6.

Proof. It is sufficient to establish inequality (14.127) for any point x of the set Q, for this inequality and inequality (14.124) immediately yield inequality (14.128) (for points x belonging to $\hat{Q}$, under the condition that Q is bounded).

We shall first prove inequality (14.127) for the case where point x is an *interior* point of the set Q. Bearing in mind that point $x^* = P_Q(x - \alpha \operatorname{grad} f(x))$ belongs to the set Q on which the function $f(x)$ is strongly convex, we shall express the value $f(x^*)$ by means of Taylor's formula with centre at point x, taking the remainder $R_2(x^*)$ in the form of Lagrange. We shall obtain

$$f(x^*) = f(x) + (\operatorname{grad} f(x), \Delta x) + \frac{1}{2} d^2 f(x + \theta \Delta x), \tag{14.129}$$

where $\Delta x = x^* - x = P_Q(x - \alpha \operatorname{grad} f(x)) - x$, $0 < \theta < 1$.

Making use of inequality (14.123) and the right-hand inequality (14.104) we obtain from Taylor's formula (14.129)

$$f(x^*) - f(x) \leqslant -\frac{1}{\alpha} |\Delta x|^2 + \frac{k_2}{2} |\Delta x|^2,$$

so that for the case of x being an interior point, inequality (14.127) has been proved.

Suppose now that x is a *boundary* point of the set Q. By the definition of the boundary point, there is a sequence $\{x_n\}$ of interior points of the set Q converging to x.

For every point x_n we obtain from Taylor's formula with centre at that point

$$f(x^*) = f(x_n) + (\operatorname{grad} f(x_n), x^* - x_n) + \frac{1}{2} d^2 f[x_n + \theta_n (x^* - x_n)], \tag{14.130}$$

where $0 < \theta_n < 1$.

Taking into account that the right-hand inequality (14.104) is valid for $d^2 f$ at any point of the set Q and that grad $f(x)$ is a continuous vector function of point x on the set Q, we find that in the limit, at $n \to \infty$, relation (14.130) yields the validity of inequality (14.127) for the boundary point x of the set Q.

We have proved Lemma 7.

We can now proceed with the proof of the fundamental theorem.

We shall first prove the fundamental theorem under the additional assumption that the closed convex set Q is also **bounded.**

We shall take an arbitrary point x_1 of the set Q and compile an iterative sequence $\{x_n\}$ of points specified by the recurrence relation (14.116) under the condition that the number α satisfies inequalities (14.115).

It immediately follows from Lemma 7, or to be more precise, from inequality (14.127), that

$$f(x_k) - f(x_{k+1}) \geqslant \left(\frac{1}{\alpha} - \frac{k_2}{2}\right) | x_k - x_{k+1} |^2 \geqslant 0.$$

Thus we see that the sequence $\{f(x_k)\}$ is nonincreasing. Since this sequence is also bounded below (by the minimum value m of the function $f(x)$ on the set Q), it is convergent (see Theorem 3.15). We designate the limit of the sequence $\{f(x_n)\}$ by μ and have $\mu \geqslant m$, where m is the minimum value of $f(x)$ on the set Q. Besides, since all the terms of a nonincreasing convergent sequence are not smaller than its limit (see Remark 3 to Theorem 3.15), the inequality

$$f(x_k) \geqslant \mu \tag{14.131}$$

is valid for all numbers k.

Let us prove that the equality $\mu = m = \min\limits_{x \in Q} f(x)$ is valid for the limit μ.

Let us assume that the equality is not valid, that is, that $\mu > m$. Then, if we denote by ν the maximum value of $f(x)$ on the set Q and by $\hat{Q}$ the subset of the points of Q for which inequalities (14.120) hold true, there will be a strictly positive constant γ, by virtue of Lemma 7, such that there will hold inequality (14.128) which leads to the following inequality

$$f(x_k) - f(x_{k+1}) \geqslant \left(\frac{1}{\alpha} - \frac{k_2}{2}\right) \gamma^2 > 0, \tag{14.132}$$

valid for any number k.

Summing up inequalities (14.132) written for the numbers k equal to 1, 2, 3, . . ., $(n-1)$, we find that for any number n

$$f(x_1) - f(x_n) \geqslant (n-1)\left(\frac{1}{\alpha} - \frac{k_2}{2}\right)\gamma^2$$

or, which is the same,

$$f(x_n) \leqslant f(x_1) - (n-1)\left(\frac{1}{\alpha} - \frac{k_2}{2}\right)\gamma^2. \tag{14.132'}$$

Inequalities (14.132'), valid for any number n, contradict inequality (14.131), since the value appearing on the right-hand side of (14.132'), becomes less than the number μ for a sufficiently large number n.

The contradiction obtained proves that the assumption $\mu > m$ is erroneous, i.e. it proves that $\mu = m$. Thus we have proved that the

sequence $\{f(x_k)\}$ converges to the minimum value m of the function $f(x)$ on the set Q.

It remains to prove that the iterative sequence $\{x_k\}$ itself converges to the point x_0 at which this minimum value can be attained*.

Let us fix an arbitrary positive number ε and denote by C_ε an open m-dimensional sphere of radius ε with centre at point x_0. Next we denote by $\hat{Q}_\varepsilon$ the part of the set Q which does not contain points of the sphere C_ε. It is clear that $\hat{Q}_\varepsilon$ is a closed bounded set, so that the function $f(x)$ attains (by virtue of Weierstrass's second theorem) its minimum value on this set of values which will be designated by m_ε.

We can assert that $m_\varepsilon > m$, since otherwise the condition of the existence of the unique point of a local minimum of the function $f(x)$ on the set Q would be violated.

We can also state that the set $\hat{Q}_\varepsilon$ contains only a finite number of points of the sequence $\{x_k\}$ (since the sequence $\{f(x_k)\}$ possessing an infinite number of terms satisfying the inequality $f(x_k) \geqslant m_\varepsilon > m$ cannot converge to the number m).

Consequently, we have proved that for any $\varepsilon > 0$ there is a number N beginning with which all the terms of the sequence $\{x_k\}$ lie in the sphere C_ε of radius ε with centre at point x_0. This precisely means that the sequence $\{x_k\}$ converges to point x_0.

Thus, the fundamental theorem has been proved for the case of a closed convex bounded set Q.

Now let Q be an **unbouded** closed convex set. Fixing again an arbitrary point x_1 of that set, we compile an iterative sequence (14.116) under the condition that the number α satisfies inequalities (14.115).

When proving the theorem on the existence of a local minimum of a strongly convex function (see 7.2), we have established that point x_0 of a local minimum of the strongly convex function $f(x)$ on the unbounded closed convex set Q lies in the part Q_R of the set Q which is contained in the sphere C_R with centre at point x_1 whose radius R has been chosen under the condition

$$\frac{\kappa_1}{2} R - |\operatorname{grad} f(x_1)| > 0.$$

It has been also established there that the subset Q_R of the set Q is a closed convex bounded set and that everywhere outside of Q_R the values $f(x)$ exceed $f(x_1)$.

Since by virtue of Lemma 7 (to be more precise, in accordance with inequality (14.127), the sequence $\{f(x_k)\}$ is nonincreasing and all the values $f(x)$ exceed $f(x_1)$ outside of Q_R, it follows that *all*

* We have already proved that the minimum value of the function $f(x)$ on the set Q can be reached at a single point of that set.

the points of the iterative sequence $\{x_k\}$ *lie within* Q_R, and, therefore, for any number k

$$P_Q(x_k - \alpha \operatorname{grad} f(x_k)) = P_{Q_R}(x_k - \alpha \operatorname{grad} f(x_k)).$$

Consequently, iterative sequence (14.116) can be replaced by

$$x_{k+1} = P_{Q_R}(x_k - \alpha \operatorname{grad} f(x_k)),$$

and then all the arguments that will follow will reduce to the bounded closed convex set Q_R, i.e. to the case we have considered above.

The proof of the fundamental theorem is complete.

Remark 1. Sequence (14.116) is especially simple in the case where the set Q coincides with the entire space E^m. Then, the equality $P_Q(x) = x$ is valid for any point x and, therefore, the recurrence formula (14.116) assumes the form

$$x_{k+1} = x_k - \alpha \operatorname{grad} f(x_k).$$

Remark 2. The method we have elucidated makes it possible, upon the satisfaction of the corresponding conditions, to seek the solution $x_0 = (\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_m)$ of the system of m functional equations:

$$\begin{cases} f_1(x) = f_1(x_1, x_2, \ldots, x_m) = 0, \\ f_2(x) = f_2(x_1, x_2, \ldots, x_m) = 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ f_m(x) = f_m(x_1, x_2, \ldots, x_m) = 0. \end{cases}$$

It is sufficient to note that the solution of the indicated system coincides with the point of a local minimum of the function

$$f(x) = f_1^2(x) + f_2^2(x) + \ldots + f_m^2(x).$$

Remark 3. The theory of seeking a local minimum of a function strongly convex (downwards) we have presented can be extended, without any complications, to the case of seeking a local maximum of a function strongly convex (upwards).

SUPPLEMENT TO CHAPTER 14

ON THE CHOICE OF THE OPTIMUM PARTITIONING OF A CLOSED INTERVAL FOR APPROXIMATE CALCULATION OF AN INTEGRAL

We have considered the case in Chapter 12 where for an approximate calculation of the integral

$$\int_a^b f(x)\,dx \tag{14.133}$$

we partitioned the closed interval $[a, b]$ into a sufficiently large number n of equal subintervals and on each of these subintervals replaced the function $f(x)$ by a polynomial of the zeroth, first or second degree. An error that appeared as a result did not take into account the specific features of $f(x)$. The question naturally arises as to the possibility of varying the points of division of the main interval $[a, b]$ and the choice for every fixed function $f(x)$ of an optimum partitioning of the main closed interval into n subintervals not equal, in general, to each other, which partitioning would ensure the minimum value of the error of the given approximate formula.

We shall present here the solution of this problem which belongs to A.N. Tikhonov and S.S. Gaisaryan*.

For an approximate calculation of integral (14.133), let us divide the interval $[a, b]$ into n subintervals by the points

$$a = x_0 < x_1 < x_2 < \ldots < x_n = b.$$

The length of the kth subinterval $[x_{k-1}, x_k]$ $(k = 1, 2, \ldots, n)$ will be denoted by h_k so that $h_k = x_k - x_{k-1}$.

Representing the integral (14.133) as the sum of the integrals

$$\int_a^b f(x)\,dx = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} f(x)\,dx, \tag{14.134}$$

we shall approximate each of the integrals on the right-hand side of (14.134) by means of one of the three approximate formulas (those of rectangles, trapezoids or parabolas). Any one of the indicated three formulas can be written in the form

$$\int_{x_{k-1}}^{x_k} f(x)\,dx = h_k \sum_{i=0}^{m-1} q_i f(x_k + t_i h_k) + R_{s,\,m,\,k}(f), \tag{14.135}$$

where the nodes t_i and the weights q_i $(i = 0, 1, \ldots, m-1)$ are chosen so that the remainder $R_{s,m,k}(f)$ would have the order h_k^s for a certain $s > 1$.**

Substituting (14.135) into (14.134), we find that

$$\int_a^b f(x)\,dx = \sum_{k=1}^{n} h_k \sum_{i=0}^{m-1} q_i f(x_k + t_i h_k) + R_{s,\,m}(f) \tag{14.136}$$

where

$$R_{s,m}(f) = \sum_{k=1}^{n} R_{s,\,m,\,k}(f). \tag{14.137}$$

Let us consider the choice of a partition $\{x_k\}$ of the interval $[a, b]$ such that the square of the error (14.137) would attain a minimum value for a fixed number of the points of division n, a fixed function $f(x)$ and a fixed approximate formula (14.135). Under these conditions the square of the error $R_{s,\,m}^2(f)$ depends only

* See A. N. Tikhonov and S. S. Gaisaryan, On optimum networks in the approximate calculation of quadrature, *USSR Computational Mathematics and Mathematical Physics*, 9, No. 5, 1969.

** In particular, for the formula of trapezoids we should put in (14.135) $m = 2$, $s = 2$, $t_0 = 0$, $t_1 = 1$, $q_0 = q_1 = \frac{1}{2}$.

on the choice of the intermediate points of division, that is, is the function of $(n-1)$ variables $x_1, x_2, \ldots, x_{n-1}$, determined in the tetrahedron $a < x_1 < x_2 < \ldots < x_{n-1} < b$. Since the tetrahedron is an open domain, the minimum of the function $R^2_{s,m}(f)$ in that tetrahedron may not, in the general case, be reached (i.e. an optimum division of the closed interval $[a, b]$ may not, in the general case, exist).

It can be proved, however, that *if the derivative* $f^{(s)}(x)$ *retains sign on the interval* $[a, b]$ *then the minimum of the square of the error* $R^2_{s,m}(f)$ *can be reached*

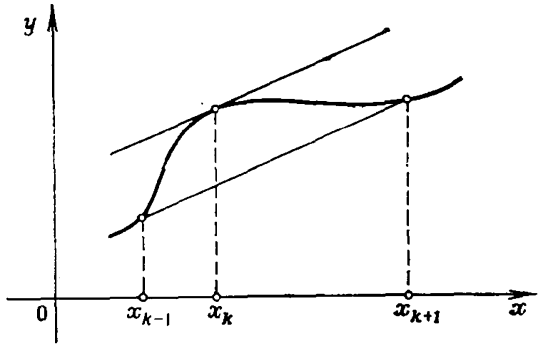


Fig. 14.4

on such a partition of the interval $[a, b]$ *whose nodes* $\bar{x}_k$ *satisfy* $(n-1)$ *equations*

$$\frac{\partial R^2_{s,m}(f)}{\partial x_k} = 0 \quad (k = 1, 2, \ldots, n-1) \tag{14.138}$$

and the additional conditions $x_0 = a$, $x_n = b$*.

We shall now treat the *formula of trapezoids* in more detail. In this case we must put in (14.136) $s = 2$, $m = 2$, $t_0 = 0$, $t_1 = 1$, $q_0 = q_1 = \frac{1}{2}$, as a result of which formula (14.136) will assume the form

$$\int_a^b f(x)\,dx = \sum_{k=1}^{n} \frac{h_k}{2}\left[f(x_k) + f(x_k + h_k)\right] + R_{2,2}(f). \tag{14.139}$$

By virtue of (14.139) and (14.134), equations (14.138) can be reduced for this case to the form

$$f(x_{k+1}) - f(x_{k-1}) = f'(x_k)(x_{k+1} - x_{k-1}) \quad (k = 1, 2, \ldots, n-1). \tag{14.140}$$

The existence of a solution for the system of equations (14.140) is ensured by the conservation of sign of the second derivative $f''(x)$ on the closed interval $[a, b]$, i.e. the conservation of the direction of the convexity of the curve $y = f(x)$ at $a \leqslant x \leqslant b$. The nodes $\{\bar{x}_k\}$ of the optimum division of the interval $[a, b]$, specified by equations (14.140), possess the following geometric property: *a secant drawn through the points of the graph of the function* $y = f(x)$ *with the abscissas* $\bar{x}_{k+1}$ and $\bar{x}_{k-1}$ *is parallel to the tangent line to the indicated graph drawn through its point with the abscissa* $\bar{x}_k$.

* The proof of this assertion can be found in Tikhonov's and Gaisaryan's paper cited above.

This property is a direct consequence of equalities (14.140) and is illustrated by Fig. 14.4 (the arguments are the same as in 8.7).

It can be proved that the nodes of the "approximate partition" $\{\widetilde{x}_k\}$ $(k = 0, 1, \ldots, n)$ specified by the recurrence equalities

$$\widetilde{x}_0 = a;\ \widetilde{x}_{k+1} = \widetilde{x}_k + \lambda \left| f''(\widetilde{x}_k) \right|^{-\frac{1}{3}} \quad (k = 0, 1, \ldots, n-1) \tag{14.141}$$

at $\lambda = \dfrac{b-a}{n}$ satisfy equations (14.140) with an error of the order of λ^3. This means that at large n the partition $\{\widetilde{x}_k\}$ is close to the optimum partition $\{\overline{x}_k\}$. Thus, for large n the computation of the integral (14.143) by the formula of trapezoids with the partition $\{\widetilde{x}_k\}$ whose nodes are successively determined by the recurrence relations (14.141), ensure the error close to the minimum value.

CHAPTER 15

THE THEORY OF IMPLICIT FUNCTIONS AND ITS APPLICATIONS

15.1. THE CONCEPT OF THE IMPLICIT FUNCTION

In mathematics and its applications we sometimes encounter problems in which a variable u, being a function of the arguments x, y, . . . by the meaning of the problem, is specified by the functional equation

$$F(u, x, y, \ldots) = 0. \tag{15.1}$$

In this case we say that as a function of the arguments x, y, . . ., u is *specified implicitly*. Thus, for instance, the function $u = -\sqrt{1 - x^2 - y^2}$ considered in the circle $x^2 + y^2 \leqslant 1$, can be specified implicitly by the functional equation

$$F(u, x, y) = u^2 + x^2 + y^2 - 1 = 0. \tag{15.2}$$

The question naturally arises as to the conditions under which the functional equation (15.1) possesses a *unique* solution with respect to u, that is, *uniquely* determines the explicit function $u = \varphi(x, y, \ldots)$. The more intricate question that also arises concerns the conditions under which this explicit function is *continuous and differentiable*. These questions are not at all simple. Thus, in the general case, the functional equation (15.2) determines in the circle $x^2 + y^2 \leqslant 1$ not only the explicit function $u = -\sqrt{1 - x^2 - y^2}$ indicated above but also infinitely many other functions. Among these other functions is the function $u = +\sqrt{1 - x^2 - y^2}$ and also any function u equal to $+\sqrt{1 - x^2 - y^2}$ for some points (x, y) of the circle $x^2 + y^2 \leqslant 1$ and to $-\sqrt{1 - x^2 - y^2}$ for all the other points of that circle. To decide on the conditions ensuring the unique solvability of equation (15.2) with respect to u, let us interpret it geometrically. Equation (15.2) determines in space (u, x, y) a sphere S of radius 1 with centre at the origin (Fig. 15.1). Let us take on the sphere S a point $M_0(\overset{\circ}{u}, \overset{\circ}{x}, \overset{\circ}{y})$ not lying in the plane Oxy, that is

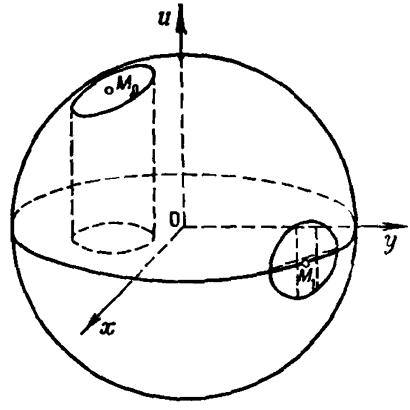


Fig. 15.1

such that $\overset{\circ}{u} \neq 0$. Evidently, the part of the sphere S belonging to a sufficiently small neighbourhood of point M_0 can be *projected in one-to-one manner on the plane Oxy*. Interpreted analytically, this means that if the function $F(u, x, y) = u^2 + x^2 + y^2 - 1$ is considered only in the indicated neighbourhood of point M_0, then equation (15.2) has a unique solution with respect to u and specifies the unique explicit function $u = +\sqrt{1 - x^2 - y^2}$ at $\overset{\circ}{u} > 0$ and $u = -\sqrt{1 - x^2 - y^2}$ at $\overset{\circ}{u} < 0$.

Now if we take a point $M_1(0, x, y)$ lying in the plane Oxy on the sphere S (Fig. 15.1), then, evidently, the part of the sphere S lying in *any* neighbourhood of M_1 *cannot be projected in one-to-one manner on the plane Oxy*. In the analytic interpretation, this means that if the function $F(u, x, y) = u^2 + x^2 + y^2 - 1$ is considered in any neighbourhood of point M_1, then equation (15.2) is not uniquely solvable with respect to u. Let us pay attention to the fact that the partial derivative $\frac{\partial F}{\partial u} = 2u$ of the function $F(u, x, y) = u^2 + x^2 + y^2 - 1$ *does not vanish at point* M_0 and vanishes at point M_1. Later on we shall establish the fact that for a unique solvability of the general functional equation (15.1) with respect to u in the neighbourhood of point M_0, of especial significance is the *nonvanishing of the partial derivative* $\frac{\partial F}{\partial u}$ *at point* M_0. At the same time we shall establish the conditions under which the explicit function which is the unique solution of equation (15.1) is *continuous and differentiable*.

In what follows we shall designate the space of the variables $(u, x, y, \ldots)$ as R and the space of the variables $(x, y, \ldots)$ as R'. To abbreviate the notation and to make geometric representation more convenient, we shall consider two variables, x and y.

15.2. THEOREM ON THE EXISTENCE AND DIFFERENTIABILITY OF AN IMPLICIT FUNCTION AND SOME OF ITS APPLICATIONS

15.2.1. Theorem on the existence and differentiability of an implicit function.

Theorem 15.1. *Suppose the function* $F(u, x, y)$ *is differentiable in some neighbourhood of point* $M_0(\overset{\circ}{u}, \overset{\circ}{x}, \overset{\circ}{y})$ *of space* R, *with the partial derivative* $\frac{\partial F}{\partial u}$ *being continuous at point* M_0. *Then, if the function* F *vanishes at point* M_0 *and the partial derivative* $\frac{\partial F}{\partial u}$ *does not vanish at that point, there can be found, for any sufficiently small positive number* ε, *a neighbourhood of point* $M'_0(\overset{\circ}{x}, \overset{\circ}{y})$ *of space* R', *such that within*

this neighbourhood there exists a single function $u = \varphi(x, y)$ *which satisfies the condition* $|u - \overset{\circ}{u}| < \varepsilon$ *and is a solution of the equation*

$$F(u, x, y) = 0, \tag{15.3}$$

the function $u = \varphi(x, y)$ *being continuous and differentiable in the indicated neighbourhood of point* M_0'.

Remark 1. The requirement of continuity of the partial derivative $\frac{\partial F}{\partial u}$ at point M_0 may be omitted in the hypothesis of Theorem 15.1, but then we should put an additional requirement that the derivative must not vanish, not only at point M_0 itself, but also in some neighbourhood of that point, and must retain a definite sign in that neighbourhood.

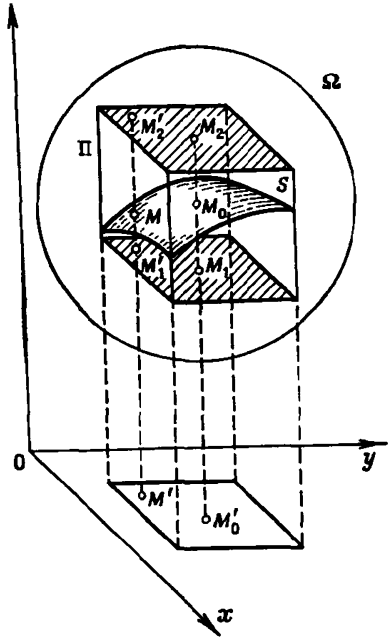


Fig. 15.2

Proof of Theorem 15.1.

1. Let us prove, first of all, that for a sufficiently small $\varepsilon > 0$ *there is a unique function* $u = \varphi(x, y)$, *in the neighbourhood of point* $M_0'(\overset{\circ}{x}, \overset{\circ}{y})$, *which satisfies the condition* $|u - \overset{\circ}{u}| < \varepsilon$ *and is a solution of equation* (15.3). To make the proof clearer, we shall present some geometric illustrations. It is known from analytic geometry that equation (15.3) specifies a certain surface S in space R (Fig. 15.2), with point M_0 lying on that surface by virtue of the condition $F(M_0) = 0$. From the point of view of geometry, the unique solvability of equation (15.3) with respect to u implies that the part of S lying in the direct proximity of point M_0 can be projected in one-to-one manner on the coordinate plane Oxy.

For the sake of definiteness, we consider the partial derivative $\frac{\partial F}{\partial u}$ to be *positive* at point M_0. Then it follows from the continuity of the indicated derivative at M_0 and from the theorem on the stability of sign of a continuous function that *there is a neighbourhood of point* M_0 throughout which $\frac{\partial F}{\partial u}$ *is everywhere positive*. We can impart to this neighbourhood the shape of a sphere Ω of a sufficiently small radius with centre at point M_0. Next we fix the positive number ε small enough for each of the points $M_1(\overset{\circ}{u} - \varepsilon, \overset{\circ}{x}, \overset{\circ}{y})$ and $M_2(\overset{\circ}{u} + \varepsilon, \overset{\circ}{x}, \overset{\circ}{y})$ to lie inside the sphere Ω (for that purpose it is sufficient

to take ε smaller than the radius of the sphere Ω). It should be emphasized that in this case ε has only zero as its lower bound and we can take it arbitrarily small. This fact will be made use of in what follows.

Let us consider a function $F(u, \overset{\circ}{x}, \overset{\circ}{y})$ of one variable on the closed interval $\overset{\circ}{u}-\varepsilon \leqslant u \leqslant \overset{\circ}{u}+\varepsilon$. From the point of view of geometry this means that we are considering a function of three variables $F(u, x, y)$ along the line segment M_1M_2 (see Fig. 15.2). The derivative $\frac{\partial F}{\partial u}(u, \overset{\circ}{x}, \overset{\circ}{y})$ being positive on the closed interval $\overset{\circ}{u}-\varepsilon \leqslant u \leqslant \overset{\circ}{u}+\varepsilon$, the function $F(u, \overset{\circ}{x}, \overset{\circ}{y})$ *increases* on that interval. But in that case, since the function is equal to zero at the middle of the indicated interval (i.e. at $u=\overset{\circ}{u}$), the function $F(u, \overset{\circ}{x}, \overset{\circ}{y})$ *possesses a negative value at the left-hand end-point and a positive value at the right-hand end-point of the indicated interval*, that is

$$F(M_1) < 0, \quad F(M_2) > 0.$$

Next we consider the functions $F(\overset{\circ}{u}-\varepsilon, x, y)$ and $F(\overset{\circ}{u}+\varepsilon, x, y)$ of two variables x and y, that is, in the geometric interpretation, we consider the function $F(u, x, y)$ in two planes parallel to the coordinate plane Oxy, the first of which passes through point M_1 and the second through point M_2. Since $F(M_1) < 0$, $F(M_2) > 0$ and the function $F(u, x, y)$ is continuous everywhere in the sphere Ω, there can be found, in accordance with the theorem on the stability of sign of a continuous function, *some neighbourhoods* of points M_1 and M_2 on the indicated planes throughout which the function F retains the same signs as at points M_1 and M_2. We can take these neighbourhoods as open squares with centres at points M_1 and M_2 and with a sufficiently small side 2δ (these squares are hatched in Fig. 15.2). The fact that the function $F(u, x, y)$ retains a constant sign on the indicated squares can be represented analytically by the inequalities

$$\left.\begin{aligned} F(\overset{\circ}{u}-\varepsilon, x, y) < 0, \\ F(\overset{\circ}{u}+\varepsilon, x, y) < 0 \end{aligned}\right\} \text{ for } |x-\overset{\circ}{x}| < \delta, \ |y-\overset{\circ}{y}| < \delta. \tag{15.4}$$

We shall make the choice of the side of the squares depend on one more condition: *we shall take* δ *small enough for the two indicated squares to lie inside the sphere* Ω (this can obviously be done, since the centres M_1 and M_2 of the squares are the interior points of the sphere Ω). With such a choice of δ, any point of the space (u, x, y) whose coordinates satisfy the inequalities

$$|x-\overset{\circ}{x}| < \delta, \quad |y-\overset{\circ}{y}| < \delta, \quad |u-\overset{\circ}{u}| < \varepsilon \tag{15.5}$$

will lie inside the sphere Ω. From the point of view of geometry, inequalities (15.5) specify an open rectangular parallelepiped with centre at point M_0 and with sides being parallel to the coordinate axes u, x, y and equal, respectively, to 2ε, 2δ and 2δ. The parallelepiped will be denoted by Π. Since parallelepiped Π lies inside the sphere Ω, *the derivative* $\frac{\partial F}{\partial u}$ *is positive everywhere inside* Π*. Furthermore, by virtue of inequalities (15.4), the function $F(u, x, y)$ *is negative on the lower base and positive on the upper base of* Π.

Let us prove now that equation (15.3) is uniquely solvable with respect to u if the function $F(u, x, y)$ is considered only for the values of u, x, y lying within the parallelepiped Π. Let us elucidate what is required to prove. Assume that $M'(x, y)$ is any point of space R' whose coordinates satisfy the inequalities

$$|x-\overset{\circ}{x}|<\delta, \quad |y-\overset{\circ}{y}|<\delta. \tag{15.6}$$

In other words, assume that $M'(x, y)$ is any point of the plane Oxy lying within the square with centre at point $M_0'(\overset{\circ}{x}, \overset{\circ}{y})$ and with the sides equal to 2δ. It is required to prove that for the coordinates x, y of point M' there is a number u from the interval $\overset{\circ}{u}-\varepsilon<u<\overset{\circ}{u}+\varepsilon$ such that $F(u, x, y)=0$ and that this number is *unique*. (Interpreted geometrically, this means that any straight line parallel to the u axis and intersecting the parallelepiped Π intersects the surface S inside Π at one and only one point.)

Having fixed the values of x and y satisfying inequalities (15.6), we consider the function $F(u, x, y)$ of the argument u on the closed interval $\overset{\circ}{u}-\varepsilon\leqslant u\leqslant\overset{\circ}{u}+\varepsilon$, that is, we consider the function $F(u, x, y)$ on the line segment $M_1'M_2'$, where M_1' and M_2' are the points of intersection of the straight line, which passes through point $M'(x, y)$ and is parallel to the Ou axis, and the bases of the parallelepiped Π (see Fig. 15.2). The derivative $\frac{\partial F}{\partial u}(u, x, y)$ being positive on the closed interval $\overset{\circ}{u}-\varepsilon\leqslant u\leqslant\overset{\circ}{u}+\varepsilon$, the function $F(u, x, y)$ increases on that interval (or, which is the same, increases on the line segment $M_1'M_2'$). But then it follows from the conditions $F(M_1')<0$, $F(M_2')>0$ that there is a single value u within the interval $\overset{\circ}{u}-\varepsilon\leqslant u\leqslant\overset{\circ}{u}+\varepsilon$ such that $F(u, x, y)=0$ (or, in terms of geometry, there is a single point M within the line segment $M_1'M_2'$ which lies on the surface S).

Suppose now that the function $u=\varphi(x, y)$ symbolizes the rule according to which every point $M'(x, y)$ from neighbourhood (15.6)

* Including the open squares lying on its bases.

is associated with a single number u from the interval $\overset{\circ}{u} - \varepsilon < u < \overset{\circ}{u} + \varepsilon$, for which $F(u, x, y) = 0$. We have proved that there is a single function $u = \varphi(x, y)$ in the neighbourhood (15.6) which satisfies the condition $|u - \overset{\circ}{u}| < \varepsilon$ and is a solution of equation (15.3).

2. Let us now prove that the *function* $u = \varphi(x, y)$ *is continuous at any point* $M'(x, y)$ *of the neighbourhood* (15.6). Since the *same conditions** *are fulfilled* for any point $M'(x, y)$ from the neighbourhood (15.6) as for point $M_0'(\overset{\circ}{x}, \overset{\circ}{y})$, it suffices to prove the continuity of the function $u = \varphi(x, y)$ *only at point* $M_0'(\overset{\circ}{x}, \overset{\circ}{y})$. We have to prove that for any sufficiently small positive ε there exists a positive number δ such that for any x and y complying with the inequalities $|x - \overset{\circ}{x}| < \delta$, $|y - \overset{\circ}{y}| < \delta$ there holds an inequality $|u - \overset{\circ}{u}| < \varepsilon$, where $u = \varphi(x, y)$, $\overset{\circ}{u} = \varphi(\overset{\circ}{x}, \overset{\circ}{y})$. If we take as ε the number we have chosen when discussing item 1, the existence of δ is ensured by inequalities (15.5). It remains to point out that in the reasoning of item 1 the positive number ε can be taken *arbitrarily small* (this was emphasized in item 1).

Thus we have established the *continuity* of the function $u = \varphi(x, y)$. Let us write the *continuity conditions* of the function $u = \varphi(x, y)$ at point $M_0'(\overset{\circ}{x}, \overset{\circ}{y})$ *in the difference form.* Denoting by Δu the total increment of the function $u = \varphi(x, y)$ at point $M_0'(\overset{\circ}{x}, \overset{\circ}{y})$ corresponding to the increments of the arguments Δx and Δy, we see that

$$\Delta u \to 0 \text{ at } \begin{cases} \Delta x \to 0, \\ \Delta y \to 0. \end{cases}$$

3. It remains to prove the *differentiability* of the function $u = \varphi(x, y)$ at any point $M'(x, y)$ of neighbourhood (15.6). By virtue of the remark made in item 2, it is sufficient to prove the differentiability of the function $u = \varphi(x, y)$ at point $M_0'(\overset{\circ}{x}, \overset{\circ}{y})$ itself. To do that, we shall calculate the total increment Δu of the function $u = \varphi(x, y)$ at point $M_0'(\overset{\circ}{x}, \overset{\circ}{y})$ corresponding to the increments of the arguments Δx and Δy. Since $F(\overset{\circ}{u}, \overset{\circ}{x}, \overset{\circ}{y}) = 0$ and $F(\overset{\circ}{u} + \Delta u, \overset{\circ}{x} + \Delta x, \overset{\circ}{y} + \Delta y) = 0$, the *total increment* ΔF *of the function* $F(u,$

* Namely, any point $M'(x, y)$ belonging to neighbourhood (15.6) is associated with point $M(u, x, y)$ of space R such that the function $F(u, x, y)$ turns into zero at point M, is differentiable in some neighbourhood of point M and has a nonzero partial derivative $\frac{\partial F}{\partial u}$ in that neighbourhood.

x, y) at point $M_0(\overset{\circ}{u}, \overset{\circ}{x}, \overset{\circ}{y})$ corresponding to the increments of the arguments Δu, Δx and Δy *is equal to zero.* By virtue of the condition of differentiability of the function $F(u, x, y)$ at point $M_0(\overset{\circ}{u}, \overset{\circ}{x}, \overset{\circ}{y})$, however, this total increment has the form

$$\Delta F = \left(\frac{\partial F}{\partial u} + \gamma\right) \Delta u + \left(\frac{\partial F}{\partial x} + \alpha\right) \Delta x + \left(\frac{\partial F}{\partial y} + \beta\right) \Delta y.$$

In this case, *all the partial derivatives* $\frac{\partial F}{\partial u}$, $\frac{\partial F}{\partial x}$ and $\frac{\partial F}{\partial y}$ *are taken at point* $M_0(\overset{\circ}{u}, \overset{\circ}{x}, \overset{\circ}{y})$; α, β and γ tend to zero as $\begin{cases} \Delta x \to 0, \\ \Delta y \to 0, \\ \Delta u \to 0. \end{cases}$

Thus we obtain

$$0 = \left(\frac{\partial F}{\partial u} + \gamma\right) \Delta u + \left(\frac{\partial F}{\partial x} + \alpha\right) \Delta x + \left(\frac{\partial F}{\partial y} + \beta\right) \Delta y. \tag{15.7}$$

In accordance with the difference form of the continuity condition of the function $u = \varphi(x, y)$ at point $M_0'(\overset{\circ}{x}, \overset{\circ}{y})$, $\Delta u \to 0$ as $\begin{cases} \Delta x \to 0, \\ \Delta y \to 0. \end{cases}$ We can, therefore, state that the condition $\begin{cases} \Delta x \to 0, \\ \Delta y \to 0 \end{cases}$ implies α, β and $\gamma \to 0$.

According to the hypothesis, the partial derivative $\frac{\partial F}{\partial u}$ *is different from zero* at point M_0. Inasmuch as $\gamma \to 0$ as $\begin{cases} \Delta x \to 0, \\ \Delta y \to 0, \end{cases}$ the expression $\frac{\partial F}{\partial u} + \gamma$ *does not vanish for sufficiently small* Δx and Δy. In that case, we can divide formula (15.7) by $\left(\frac{\partial F}{\partial u} + \gamma\right)$ and obtain as a result

$$\Delta u = \left(-\frac{\frac{\partial F}{\partial x} + \alpha}{\frac{\partial F}{\partial u} + \gamma}\right) \Delta x + \left(-\frac{\frac{\partial F}{\partial u} + \beta}{\frac{\partial F}{\partial u} + \gamma}\right) \Delta y. \tag{15.8}$$

Proceeding from the theorem on the limiting value of the quotient of two functions we can state that

$$-\frac{\frac{\partial F}{\partial x} + \alpha}{\frac{\partial F}{\partial u} + \gamma} = -\frac{\frac{\partial F}{\partial x}}{\frac{\partial F}{\partial u}} + \mu, \quad -\frac{\frac{\partial F}{\partial y} + \beta}{\frac{\partial F}{\partial u} + \gamma} = \frac{\frac{\partial F}{\partial y}}{\frac{\partial F}{\partial u}} + \nu, \tag{15.9}$$

where μ and $\nu \to 0$ as $\begin{cases} \Delta x \to 0, \\ \Delta y \to 0. \end{cases}$

Comparing formulas (15.8) and (15.9), we finally obtain

$$\Delta u = \left(-\frac{\frac{\partial F}{\partial x}}{\frac{\partial F}{\partial u}}\right)\Delta x + \left(-\frac{\frac{\partial F}{\partial y}}{\frac{\partial F}{\partial u}}\right)\Delta y + \mu\,\Delta x + \nu\,\Delta y. \tag{15.10}$$

Formula (15.10) proves the differentiability of the function $u = \varphi(x, y)$ at point $M_0'(\overset{\circ}{x}, \overset{\circ}{y})$. The proof of Theorem 15.1 is complete.

Remark 2. The proof we have carried out can be extended, without difficulty, to the case of an implicit function, depending not on two but on any finite number of the arguments $x_1, x_2, \ldots, x_m$*. We have taken the case of two arguments x and y only because it admits of a visual geometric representation in the space (u, x, y).

15.2.2. Calculating partial derivatives of a function specified implicitly. We shall now busy ourselves with the calculation of partial derivatives of a function specified implicitly by equation (15.3). Suppose the conditions of Theorem 15.1 have been satisfied. Then, representation (15.10) is valid for the total increment of the function $u = \varphi(x, y)$. This representation and Theorem 14.9 enable us to assert that the partial derivatives of the function $u = \varphi(x, y)$ are determined by the formulas

$$\frac{\partial u}{\partial x} = -\frac{\frac{\partial F}{\partial x}}{\frac{\partial F}{\partial u}}, \quad \frac{\partial u}{\partial y} = -\frac{\frac{\partial F}{\partial y}}{\frac{\partial F}{\partial u}}. \tag{15.11}$$

Similar formulas are also valid in the case where an implicitly specified function depends not on two but on any finite number of arguments $x_1, x_2, \ldots, x_m$. In this case

$$\frac{\partial u}{\partial x_k} = -\frac{\frac{\partial F}{\partial x_k}}{\frac{\partial F}{\partial u}} \quad (k = 1, 2, \ldots, m).$$

If we want to ensure the existence of the *second-order* partial derivatives of the implicitly specified function $u = \varphi(x, y)$, we must naturally strengthen the requirements imposed on the function $F(u, x, y)$ in Theorem 15.1, namely, we must additionally require that the function $F(u, x, y)$ be twice differentiable at the point in question. Under these assumptions, we shall now discuss the *calculation of the second-order partial derivatives.*

It is useful to introduce at this stage the concept of a total partial derivative of a function that we shall need later on. Let us assume

* It can depend on one argument, in a particular case.

that we are given a differentiable function of three arguments $\Phi(u, x, y)$, one of the arguments, u, being itself a differentiable function of the other two arguments x and y. Then, the function $\Phi(u, x, y)$ can be regarded as a composite function of x, y. The partial derivatives of this composite function with respect to x and y will be called *total partial derivatives of the function* $\Phi(u, x, y)$ *with respect to* x *and* y and denoted as $D\Phi/Dx$ and $D\Phi/Dy$. By the rule of differentiation of a composite function we shall obtain the following formulas for the indicated total partial derivatives:

$$\frac{D\Phi}{Dx}=\frac{\partial\Phi}{\partial u}\frac{\partial u}{\partial x}+\frac{\partial\Phi}{\partial x}, \quad \frac{D\Phi}{Dy}=\frac{\partial\Phi}{\partial u}\frac{\partial u}{\partial y}+\frac{\partial\Phi}{\partial y}.$$

Let us try to carry out calculations of the second-order partial derivatives of an implicitly specified function. For the sake of definiteness, we shall calculate the derivative $\partial^2 u/\partial y\partial x$. Differentiating the first formula (15.11) with respect to y and taking into consideration that each of the partial derivatives $\partial F/\partial x$ and $\partial F/\partial u$ depends on three arguments u, x, y, the first of which is itself a function of x and y, we obtain

$$\frac{\partial^2 u}{\partial y\,\partial x}=\frac{D\left[-\dfrac{\dfrac{\partial F}{\partial x}}{\dfrac{\partial F}{\partial u}}\right]}{Dy}=\frac{-\dfrac{\partial F}{\partial u}\dfrac{D\left[\dfrac{\partial F}{\partial x}\right]}{Dy}+\dfrac{\partial F}{\partial x}\dfrac{D\left[\dfrac{\partial F}{\partial u}\right]}{Dy}}{\left(\dfrac{\partial F}{\partial u}\right)^2}=$$

$$=\frac{-\dfrac{\partial F}{\partial u}\left(\dfrac{\partial^2 F}{\partial x\,\partial u}\dfrac{\partial u}{\partial y}+\dfrac{\partial^2 F}{\partial x\,\partial y}\right)+\dfrac{\partial F}{\partial x}\left(\dfrac{\partial^2 F}{\partial u^2}\dfrac{\partial u}{\partial y}+\dfrac{\partial^2 F}{\partial u\,\partial y}\right)}{\left(\dfrac{\partial F}{\partial u}\right)^2}.$$

Substituting into the formulas obtained the expression $\partial u/\partial y$ determined by the second formula (15.11), we finally have

$$\frac{\partial^2 u}{\partial y\,\partial x}=$$

$$=\frac{\dfrac{\partial^2 F}{\partial x\,\partial u}\dfrac{\partial F}{\partial y}\dfrac{\partial F}{\partial u}-\dfrac{\partial^2 F}{\partial x\,\partial y}\left(\dfrac{\partial F}{\partial u}\right)^2-\dfrac{\partial^2 F}{\partial u^2}\dfrac{\partial F}{\partial x}\dfrac{\partial F}{\partial y}+\dfrac{\partial^2 F}{\partial y\,\partial u}\dfrac{\partial F}{\partial x}\dfrac{\partial F}{\partial u}}{\left(\dfrac{\partial F}{\partial u}\right)^3}. \qquad (15.12)$$

The partial derivatives $\partial^2 u/\partial x^2$ and $\partial^2 u/\partial y^2$ can be calculated quite analogously. A similar method can be used to calculate partial derivatives of the third and subsequent orders*.

* Under the condition that the function $F(u, x, y)$ is differentiable at the given point a requisite number of times.

Examples. (1) Calculate the partial derivative $\partial^2 u/\partial y \partial x$ of the function $u = \varphi(x, y)$ specified by the equation

$$x + y + u - e^{-(x+y+u)} = 0.$$

First of all, using formulas (15.11), we shall calculate the following first-order derivatives:

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial y} = -\frac{1+e^{-(x+y+u)}}{1+e^{-(x+y+u)}} = -1.$$

Then it is evident that $\frac{\partial^2 u}{\partial y \partial x} = 0$.

(2) The same problem for the function specified by the equation

$$u^2 + x^2 + y^2 - a^2 = 0.$$

Using formulas (15.11), we obtain

$$\frac{\partial u}{\partial x} = -\frac{x}{u}, \quad \frac{\partial u}{\partial y} = -\frac{y}{u}.$$

Then we have

$$\frac{\partial^2 u}{\partial y \partial x} = \frac{D\left(\frac{\partial u}{\partial x}\right)}{Dy} = \frac{D\left(-\frac{x}{u}\right)}{Dy} = \frac{x\frac{\partial u}{\partial y}}{u^2} = -\frac{xy}{u^3}.$$

15.2.3. Singularities of a surface and of a plane curve. Let us consider a certain surface S (a plane curve L) specified by the equation $F(x, y, z) = 0$ ($F(x, y) = 0$) in a given rectangular Cartesian system of coordinates. As concerns the function $F(x, y, z)$ ($F(x, y)$), we shall assume that it possesses first-order continuous partial derivatives with respect to all the arguments everywhere in some neighbourhood of any point of surface S (curve L). We shall call the given point of surface S (curve L) a singular point (a singularity) if all the first-order partial derivatives of the function

$$F(x, y, z) \quad (F(x, y))$$

vanish at that point. Theorem 15.1 cannot be applied to the equation $F(x, y, z) = 0$ ($F(x, y) = 0$) in the neighbourhood of a singular point, that is, it cannot be asserted that the equation is solvable at least with respect to one of the variables x, y, z (x, y). Thus, the part of surface S (curve L) adjacent to a singular point may not admit of a one-to-one projection on any of the coordinate planes (on any one of the coordinate axes). The structure of surface S (curve L) in the neighbourhood of a singular point may be very intricate and requires additional investigation.

Points of surface S (curve L) which are not singular are usually called *regular* (*nonsingular*) points. Theorem 15.1 works in the neighbourhood of a regular point, and so the part of surface S (curve L) adjacent to a regular point admits of one-to-one projection on at

least one of the coordinate planes (on at least one of the coordinate axes), and this substantially simplifies the investigation of that part of the surface.

Examples. (1) Find the singularities of the circular cone $x^2+y^2-z^2=0$. Since $F(x, y, z)=x^2+y^2-z^2$, it follows that $\frac{\partial F}{\partial x}=2x$, $\frac{\partial F}{\partial y}=2y$, $\frac{\partial F}{\partial z}=-2z$. The only singularity is the origin. It is well known that in the neighbourhood of that point the surface of the cone cannot be projected in a one-to-one

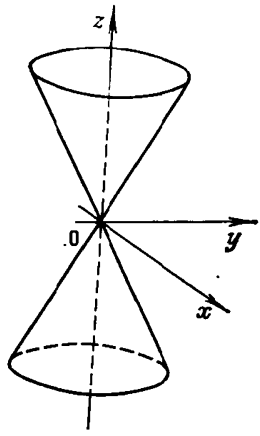


Fig. 15.3

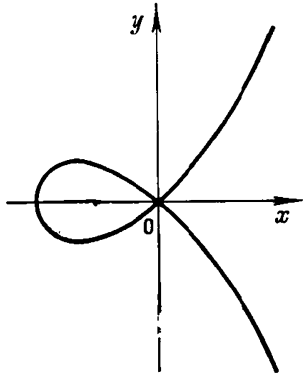


Fig. 15.4

manner on any one of the coordinate planes (Fig. 15.3).

(2) The same problem for the plane curve $x^2-y^2+x^3=0$. The partial derivatives have the form $\frac{\partial F}{\partial x}=2x+3x^2=x(2+3x)$, $\frac{\partial F}{\partial y}=-2y$. Both partial derivatives vanish at two points of the plane, (0, 0) and $\left(-\frac{2}{3}, 0\right)$. Only the first of these points belongs to the curve being considered, i.e. is singular. Having constructed the curve $x^2-y^2+x^3=0$ in the neighbourhood of point (0, 0), we shall verify that this point is a point of self-intersection of the graph (Fig. 15.4). It is clear that in the neighbourhood of this point the curve cannot be projected in a one-to-one manner either on the Ox axis or on the Oy axis.

15.2.4. Conditions ensuring the existence of an inverse of the function $y=f(x)$. Let us apply Theorem 15.1 to elucidate the conditions under which the function $y=f(x)$ possesses in some neighbourhood of point x_0 an *inverse function* $x=f^{-1}(y)$ defined in a certain neighbourhood of point y_0, where $y_0=f(x_0)$. Let us consider $y=$

$= f(x)$ to be a function specified by a functional equation of the form $F(x, y) = f(x) - y = 0$.

In that case the problem of the existence of an inverse coincides with the problem concerning the solvability of the functional equation with respect to x. As a consequence of Theorem 15.1 and Remark 1 following the formulation of the theorem, we obtain the assertion: *if the function $y = f(x)$ possesses a nonzero derivative in some neighbourhood of point x_0, then there exists an inverse $x = f^{-1}(y)$ of that function in the neighbourhood of x_0 determined and differentiable in some neighbourhood of point y_0, where $y_0 = f(x_0)$. The derivative of the inverse function at point y_0 is equal to $\frac{1}{f'(x_0)}$ by virtue of the second formula* (15.11).

15.3. IMPLICIT FUNCTIONS SPECIFIED BY A SYSTEM OF FUNCTIONAL EQUATIONS

15.3.1. Theorem on the solvability of a system of functional equations. In the previous section we discussed the problem of the existence and differentiability of an implicit function specified by means of *one* functional equation. Here we shall consider a similar problem concerning *a set of m* (m being any natural number) *implicit functions specified by a system of functional equations.*

Thus, let us suppose that m functions

$$\left.\begin{aligned} u_1 &= \varphi_1(x_1, x_2, \ldots, x_n), \\ u_2 &= \varphi_2(x_1, x_2, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots \\ u_m &= \varphi_m(x_1, x_2, \ldots, x_n) \end{aligned}\right\} \tag{15.13}$$

are being sought as a solution of a system of m functional equations

$$\left.\begin{aligned} F_1(u_1, u_2, \ldots, u_m, x_1, x_2, \ldots, x_n) &= 0, \\ F_2(u_1, u_2, \ldots, u_m, x_1, x_2, \ldots, x_n) &= 0, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ F_m(u_1, u_2, \ldots, u_m, x_1, x_2, \ldots, x_n) &= 0. \end{aligned}\right\} \tag{15.14}$$

Let us take the problem of solvability of the system of functional equations (15.14) with respect to $u_1, u_2, \ldots, u_m$. The term "a solution of system (15.14)" will mean, in what follows, a set of m functions (15.13) such that when substituted into system (15.14) they turn all the equations of the system into identities. This solution will be called continuous and differentiable in a certain domain D of variation of the variables $x_1, x_2, \ldots, x_n$, if each of the functions (15.13) is continuous and differentiable in the domain D. Let us agree to use the symbol R for the space of $(m + n)$ variables u_1,

$u_2, \ldots, u_m, x_1, x_2, \ldots, x_n$ and the symbol R' for the space of n variables $x_1, x_2, \ldots, x_n$.

We shall consider m functions $F_1, F_2, \ldots, F_m$ entering into the left-hand sides of system (15.14) and form the following determinant from the partial derivatives of these functions:

$$\Delta = \begin{vmatrix} \frac{\partial F_1}{\partial u_1} & \frac{\partial F_1}{\partial u_2} & \cdots & \frac{\partial F_1}{\partial u_m} \\ \frac{\partial F_2}{\partial u_1} & \frac{\partial F_2}{\partial u_2} & \cdots & \frac{\partial F_2}{\partial u_m} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial F_m}{\partial u_1} & \frac{\partial F_m}{\partial u_2} & \cdots & \frac{\partial F_m}{\partial u_m} \end{vmatrix}. \tag{15.15}$$

The determinant of form (15.15) is known as *Jacobi's* determinant* (or simply the *Jacobian*) of the functions $F_1, F_2, \ldots, F_m$ with respect to the variables $u_1, u_2, \ldots, u_m$ and is usually designated briefly as

$$\frac{D(F_1, F_2, \ldots, F_m)}{D(u_1, u_2, \ldots, u_m)}.$$

The following remarkable assertion holds true.

Theorem 15.2 (generalization of Theorem 15.1). *Let m functions*

$$\left.\begin{aligned} &F_1(u_1, u_2, \ldots, u_m, x_1, \ldots, x_n), \\ &F_2(u_1, u_2, \ldots, u_m, x_1, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &F_m(u_1, u_2, \ldots, u_m, x_1, \ldots, x_n) \end{aligned}\right\} \tag{15.16}$$

be differentiable in some neighbourhood of point $M_0(\overset{\circ}{u}_1, \overset{\circ}{u}_2, \ldots, \overset{\circ}{u}_m, \overset{\circ}{x}_1, \ldots, \overset{\circ}{x}_n)$ of space R, the partial derivatives of the functions with respect to the variables $u_1, u_2, \ldots, u_m$ being continuous at point M_0. In that case, if all the functions (15.16) *vanish at point M_0 and the Jacobian $\frac{D(F_1, F_2, \ldots, F_m)}{D(u_1, u_2, \ldots, u_m)}$ is different from zero, then, for sufficiently small positive numbers $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$, there is a neighbourhood of point $M_0'(\overset{\circ}{x}_1, \ldots, \overset{\circ}{x}_n)$ of space R' such that there exist, within this neighbourhood, m and only m functions* (15.13) *which satisfy the conditions $|u_1 - \overset{\circ}{u}_1| < \varepsilon_1$, $|u_2 - \overset{\circ}{u}_2| < \varepsilon_2, \ldots, |u_m - \overset{\circ}{u}_m| < \varepsilon_m$ and are a solution of the system of equations* (15.14), *the solution being continuous and differentiable in the indicated neighbourhood of point M_0'.*

* Karl Gustav Jacob Jacobi (1804-1851), a German mathematician.

Remark. At $m = 1$ Theorem 15.2 passes into Theorem 15.1 proved above* because then Jacobian (15.15) turns into the partial derivative $\partial F_1/\partial u_1$.

The proof of Theorem 15.2 will be carried out by the method of mathematical induction. We have proved the theorem for $m = 1$. Therefore, it is sufficient to assume Theorem 15.2 to be valid for the system of $m - 1$ functional equations and prove the validity of the theorem for the system of m functional equations as well. Since by the assumption the Jacobian

$$\Delta = \frac{D(F_1, F_2, \ldots, F_m)}{D(u_1, u_2, \ldots, u_m)} = \begin{vmatrix} \dfrac{\partial F_1}{\partial u_1} & \cdots & \dfrac{\partial F_1}{\partial u_{m-1}} & \dfrac{\partial F_1}{\partial u_m} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{\partial F_{m-1}}{\partial u_1} & \cdots & \dfrac{\partial F_{m-1}}{\partial u_{m-1}} & \dfrac{\partial F_{m-1}}{\partial u_m} \\ \dfrac{\partial F_m}{\partial u_1} & \cdots & \dfrac{\partial F_m}{\partial u_{m-1}} & \dfrac{\partial F_m}{\partial u_m} \end{vmatrix} \tag{15.17}$$

is nonzero at point M_0, *at least one of the minors* of the $(m - 1)$th order** of the Jacobian is nonzero at point M_0. Without violating generality, we shall assume that different from zero at point M_0 is the minor enclosed in a dotted line in the upper left-hand corner. Then, by the induction assumption, the first $m - 1$ equations of system (15.14) are solvable with respect to $u_1, u_2, \ldots, u_{m-1}$. To be more precise, for sufficiently small positive numbers $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{m-1}$ there is a neighbourhood of point $M_0''(\overset{\circ}{u}_m, \overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_n)$ of space R'' of the variables $(u_m, x_1, x_2, \ldots, x_n)$ such that within this neighbourhood $m - 1$ functions

$$\left.\begin{aligned} u_1 &= \Phi_1(u_m, x_1, x_2, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ u_{m-1} &= \Phi_{m-1}(u_m, x_1, x_2, \ldots, x_n) \end{aligned}\right\} \tag{15.18}$$

are determined which satisfy the conditions $|u_1 - \overset{\circ}{u}_1| < \varepsilon_1, \ldots, |u_{m-1} - \overset{\circ}{u}_{m-1}| < \varepsilon_{m-1}$ and are, under these conditions, the unique continuous and differentiable solution of the system of the first $m - 1$ equations (15.14).

Let us substitute functions (15.18) we have obtained into the left-hand side of the last equation (15.14). This will turn the left-hand side of the last equation (15.14) into a function depen-

---

* We should also take into account Remark 2 to Theorem 15.1.

** Recall that the minor of the $(m-1)$th order of a given determinant of the mth order is the determinant of order $(m - 1)$ obtained from the given mth order determinant by means of deleting one row and one column.

dent only on $u_m, x_1, \ldots, x_n$

$$F_m(u_1, \ldots, u_{m-1}, u_m, x_1, \ldots, x_n) = F_m[\Phi_1(u_m, x_1, \ldots, x_n), \ldots, \Phi_{m-1}(u_m, x_1, \ldots, x_n), u_m, x_1, \ldots, x_n] = \Psi(u_m, x_1, \ldots, x_n) \tag{15.19}$$

(we have designated this function as Ψ). Thus, the last equation of system (15.14) leads to the equation

$$\Psi(u_m, x_1, \ldots, x_n) = 0. \tag{15.20}$$

By virtue of equality (15.19), $\Psi(u_m, x_1, \ldots, x_n)$ can be regarded as a composite function of its arguments. Then, applying the theorem on the differentiability of a composite function, we can state that the function $\Psi(u_m, x_1, \ldots, x_n)$ *is differentiable in some neighbourhood of point* $M_0''(\overset{\circ}{u}_m, \overset{\circ}{x}_1, \ldots, \overset{\circ}{x}_n)$ *of space* R''. Equality (15.19) and the last equation (15.14) permit us to assert that $\Psi(\overset{\circ}{u}_m, \overset{\circ}{x}_1, \ldots, \overset{\circ}{x}_n) = 0$. Consequently, to prove that Theorem 15.1 is applicable to equation (15.20) and that the equation is solvable with respect to u_m, it is sufficient to establish that the partial derivative $\partial\Psi/\partial u_m$ is continuous and different from zero at point M_0''. To establish this fact, let us calculate the indicated partial derivative. Let us substitute functions (15.18) which are the solution of system (15.14) into the first $m-1$ equations of this system and then differentiate the identities obtained with respect to u_m. We shall have

$$\frac{\partial F_1}{\partial u_1}\frac{\partial \Phi_1}{\partial u_m} + \ldots + \frac{\partial F_1}{\partial u_{m-1}}\frac{\partial \Phi_{m-1}}{\partial u_m} + \frac{\partial F_1}{\partial u_m} = 0, \tag{15.21^1}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\frac{\partial F_{m-1}}{\partial u_1}\frac{\partial \Phi_1}{\partial u_m} + \ldots + \frac{\partial F_{m-1}}{\partial u_{m-1}}\frac{\partial \Phi_{m-1}}{\partial u_m} + \frac{\partial F_{m-1}}{\partial u_m} = 0. \tag{15.21^{m-1}}$$

Next we shall differentiate equality (15.19) with respect to u_m. We obtain

$$\frac{\partial F_m}{\partial u_1}\frac{\partial \Phi_1}{\partial u_m} + \ldots + \frac{\partial F_m}{\partial u_{m-1}}\frac{\partial \Phi_{m-1}}{\partial u_m} + \frac{\partial F_m}{\partial u_m} = \frac{\partial \Psi}{\partial u_m}. \tag{15.21^m}$$

Multiplying now equalities (15.21^1)-(15.21^m) by the respective algebraic cofactors $\Delta_1, \Delta_2, \ldots, \Delta_m$ of the elements of the last column of Jacobian (15.17) and adding up the results, we obtain

$$\sum_{k=1}^{m-1}\frac{\partial \Phi_k}{\partial u_m}\left[\Delta_1\frac{\partial F_1}{\partial u_k} + \Delta_2\frac{\partial F_2}{\partial u_k} + \ldots + \Delta_m\frac{\partial F_m}{\partial u_k}\right] +$$

$$+\left(\Delta_1\frac{\partial F_1}{\partial u_m} + \Delta_2\frac{\partial F_2}{\partial u_m} + \ldots + \Delta_m\frac{\partial F_m}{\partial u_m}\right) = \Delta_m\frac{\partial \Psi}{\partial u_m}.$$

Since the sum of the products of elements of a given column of the determinant by the respective algebraic cofactors of elements of

this (other) column is equal to the determinant (to zero), each expression in brackets is equal to zero and each expression in parentheses to Jacobian (15.17).

Thus we obtain

$$\Delta = \Delta_m \frac{\partial \Psi}{\partial u_m}. \tag{15.22}$$

Here Δ denotes Jacobian (15.17) and Δ_m is the algebraic cofactor of the last element of the last column which coincides with the minor enclosed in a dotted line and, by the assumption, *different from zero* at point M_0. Having divided equality (15.22) by Δ_m, we finally have

$$\frac{\partial \Psi}{\partial u_m} = \frac{\Delta}{\Delta_m}. \tag{15.23}$$

Formula (15.23) valid at point M_0'' proves the continuity of the partial derivative $\partial \Psi / \partial u_m$ at point M_0'' (because Δ and Δ_m consist of the partial derivatives of functions (15.16) with respect to $u_1, u_2, \ldots, u_m$ which are continuous at point M_0). Besides, formula (15.23) implies that $\partial \Psi / \partial u_m$ is nonzero at point M_0'' (since the Jacobian Δ is nonzero at point M_0). We have thus proved that Theorem 15.1 can be applied to equation (15.20).

In accordance with this theorem, for a sufficiently small positive number ε_m there is a neighbourhood of point $M_0'(\overset{\circ}{x}_1, \ldots, \overset{\circ}{x}_n)$ of space R' such that everywhere within this neighbourhood the function

$$u_m = \varphi_m (x_1, x_2, \ldots, x_n) \tag{15.24}$$

is defined which satisfies the condition $|u_m - \overset{\circ}{u}_m| < \varepsilon_m$ and, under this condition, is the unique continuous and differentiable solution of equation (15.20). Bearing in mind that functions (15.18) are solutions of the first $m - 1$ equations (15.14) *for any* $u_m, x_1, \ldots, x_n$ from the neighbourhood of point M_0'' and substituting function (15.24) we have obtained into (15.18), we shall have functions dependent only on the variables $x_1, \ldots, x_n$:

$$u_1 = \Phi_1 [\varphi_m (x_1, \ldots, x_n), x_1, \ldots, x_n] = \varphi_1 (x_1, \ldots, x_n),$$

$$\ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots$$

$$u_{m-1} = \Phi_{m-1} [\varphi_m (x_1, \ldots, x_n), x_1, \ldots, x_n] = \varphi_{m-1} (x_1, \ldots, x_n).$$

(We have designated these functions as $\varphi_1, \ldots, \varphi_{m-1}$.) The theorem on differentiability of a composite function enables us to assert that each of the functions $\varphi_1, \ldots, \varphi_{m-1}$ is differentiable in the neighbourhood of point $M_0'(\overset{\circ}{x}_1, \ldots, \overset{\circ}{x}_n)$. Thus we have proved

for sure that the m functions

$$\left.\begin{aligned} u_1 &= \varphi_1(x_1, \ldots, x_n), \\ u_2 &= \varphi_2(x_1, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots \\ u_m &= \varphi_m(x_1, \ldots, x_n) \end{aligned}\right\} \tag{15.25}$$

satisfy the conditions $|u_1 - \overset{\circ}{u}_1| < \varepsilon_1, \ldots, |u_m - \overset{\circ}{u}_m| < \varepsilon_m$ in the neighbourhood of point M_0' and under these conditions are the unique continuous and differentiable solution of system (15.14) in some neighbourhood of point $M_0'(\overset{\circ}{x}_1, \ldots, \overset{\circ}{x}_n)$.

It remains to prove that functions (15.25) are the unique solution of system (15.14) satisfying the conditions $|u_1 - \overset{\circ}{u}_1| < \varepsilon_1, \ldots$ $\ldots, |u_m - \overset{\circ}{u}_m| < \varepsilon_m$ (for sufficiently small positive $\varepsilon_1, \ldots, \varepsilon_m$).

Let us assume that besides functions (15.25) there also exist m functions

$$\left.\begin{aligned} \hat{u}_1 &= \hat{\varphi}_1(x_1, x_2, \ldots, x_n), \\ \hat{u}_2 &= \hat{\varphi}_2(x_1, x_2, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots \\ \hat{u}_m &= \hat{\varphi}_m(x_1, x_2, \ldots, x_n), \end{aligned}\right\} \tag{15.25'}$$

which are also a solution of system (15.14) and satisfy the conditions $|\hat{u}_1 - \overset{\circ}{u}_1| < \varepsilon_1, \ldots, |\hat{u}_m - \overset{\circ}{u}_m| < \varepsilon_m$.

Then, by virtue of the induction assumption, the first $m - 1$ functions (15.25) are, for a given $u_m = \hat{u}_m$, the unique and differentiable solution of the system of the first $m - 1$ equations (15.14). But, for the given u_m, the unique solution of the system of the first $m - 1$ equations (15.14) is ensured by equalities (15.18). Thus, the relationships

$$\left.\begin{aligned} \hat{u}_1 &= \Phi_1(\hat{u}_m, x_1, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots \\ \hat{u}_{m-1} &= \Phi_{m-1}(\hat{u}_m, x_1, \ldots, x_n) \end{aligned}\right\} \tag{15.18'}$$

hold true, in which $\Phi_1, \ldots, \Phi_{m-1}$ are the same functions as in (15.18).

In that case, the last equation (15.14) and relationships (15.19) allow us to assert that $\hat{u}_m$ is the unique solution of equation (15.20), i.e. $\hat{u}_m = u_m$.

If the equality $\hat{u}_m = u_m$ is valid, relationships (15.18') and (15.18) immediately yield $\hat{u}_1 = u_1, \ldots, \hat{u}_{m-1} = u_{m-1}$.

The proof of Theorem 15.2 is complete.

15.3.2. Calculating partial derivatives of functions specified implicitly by a system of functional equations. We shall assume here that the conditions of Theorem 15.2 are satisfied and shall calculate the partial derivatives of functions (15.25). We shall substitute functions (15.25) into the system of equations (15.14) whose solution they are and differentiate the resulting identities with respect to x_l ($l = 1, 2, \ldots, n$). We shall have

$$\left.\begin{aligned} &\frac{\partial F_1}{\partial u_1}\frac{\partial u_1}{\partial x_l}+\ldots+\frac{\partial F_1}{\partial u_m}\frac{\partial u_m}{\partial x_l}+\frac{\partial F_1}{\partial x_l}=0,\\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ &\frac{\partial F_m}{\partial u_1}\frac{\partial u_1}{\partial x_l}+\ldots+\frac{\partial F_m}{\partial u_m}\frac{\partial u_m}{\partial x_l}+\frac{\partial F_m}{\partial x_l}=0. \end{aligned}\right\} \tag{15.26}$$

Equations (15.26) are a linear system of equations with respect to m variables $\frac{\partial u_1}{\partial x_l}, \ldots, \frac{\partial u_m}{\partial x_l}$. The determinant of the system, Jacobian (15.17), is different from zero in the neighbourhood of point M_0. Consequently, system (15.26) has a unique solution given by Cramer's formulas

$$\frac{\partial u_k}{\partial x_l}=\frac{\dfrac{D(F_1, F_2, \ldots, F_m)}{D(u_1, \ldots, u_{k-1}, x_l, u_{k+1}, \ldots, u_m)}}{\dfrac{D(F_1, F_2, \ldots, F_m)}{D(u_1, u_2, \ldots, u_m)}}.$$

The expressions for the partial derivatives of the second and subsequent orders* can be obtained by differentiating these formulas.

15.3.3. One-to-one mapping of two sets of an m-dimensional space. *Let us consider m functions*

$$\left.\begin{aligned} u_1&=\varphi_1(x_1, \ldots, x_m),\\ u_2&=\varphi_2(x_1, \ldots, x_m),\\ &\cdots\cdots\cdots\\ u_m&=\varphi_m(x_1, \ldots, x_m) \end{aligned}\right\} \tag{15.27}$$

in some neighbourhood of point $M_0(\overset{\circ}{x}_1, \ldots, \overset{\circ}{x}_m)$. These m functions map the neighbourhood of point M_0 on a certain set $\{N\}$ of the m-dimensional space of the variables $u_1, u_2, \ldots, u_m$. This mapping is said to be *one-to-one* if every point from the indicated neighbourhood of point M_0 is associated with only one point of the set $\{N\}$, so that every point of the set $\{N\}$ corresponds to only one point belonging to the neighbourhood of point M_0.

Theorem 15.2 yields the following direct assertion.

If functions (15.27) *are differentiable in the neighbourhood of point M_0, with all the first-order partial derivatives being continuous at point M_0 itself and with the Jacobian $\frac{D(\varphi_1, \ldots, \varphi_m)}{D(x_1, \ldots, x_m)}$ being different from zero at that point, then functions* (15.27) *perform a one-to-one mapping of a certain neighbourhood of point $M_0(\overset{\circ}{x}_1, \ldots, \overset{\circ}{x}_m)$ on a certain neighbourhood of point $N_0(\overset{\circ}{u}_1, \ldots, \overset{\circ}{u}_m)$, where $\overset{\circ}{u}_i = \varphi_i(\overset{\circ}{x}_1, \ldots, \overset{\circ}{x}_m)$ ($i = 1, 2, \ldots, m$).*

Indeed, relationships (15.27) may be regarded as a system of functional equations with respect to $x_1, x_2, \ldots, x_m$, for which the conditions of Theorem 15.2 are satisfied. But then everywhere in some neighbourhood of point

* The existence of these partial derivatives is ensured by the additional restrictions imposed on functions (15.16).

$N_0(\overset{\circ}{u}_1, \ldots, \overset{\circ}{u}_m)$ this system possesses the unique solution

$$\left.\begin{aligned} x_1 &= \psi_1(u_1 \ldots, u_m), \\ &\cdots\cdots\cdots\cdots \\ x_m &= \psi_m(u_1 \ldots, u_m). \end{aligned}\right\} \tag{15.27'}$$

It is evident that functions (15.27') perform an inverse mapping.

Note that under the conditions of the assertion we have formulated, both functions (15.27) performing a direct mapping and functions (15.27') performing in inverse mapping are *continuous*. A one-to-one mapping possessing this property is called *homeomorphic*.

15.4. FUNCTIONAL DEPENDENCE

15.4.1. The concept of functional dependence. A sufficient condition of independence. Let m functions of the same n variables

$$\left.\begin{aligned} u_1 &= \varphi_1(x_1, x_2, \ldots, x_n), \\ u_2 &= \varphi_2(x_1, x_2, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots\cdots \\ u_m &= \varphi_m(x_1, x_2, \ldots, x_n) \end{aligned}\right\} \tag{15.28}$$

be defined and differentiable in some open n-dimensional domain* D.

We shall say that *one of these functions*, say u_k, *depends on other functions in the domain D if for all the points $(x_1, x_2, \ldots, x_n)$ of the domain D at once*

$$u_k = \Phi(u_1, \ldots, u_{k-1}, u_{k+1}, \ldots, u_m), \tag{15.29}$$

where Φ is a certain function defined and differentiable in the corresponding domain of its arguments. Functions $u_1, u_2, \ldots, u_m$ are said to be *dependent in the domain D* if one of them (no matter which) depends on the rest of the functions in the domain D.

Now if there is no differentiable function Φ such that an identity of the form (15.29) is valid for all the points of the domain D simultaneously, we shall call the functions $u_1, u_2, \ldots, u_m$ independent in the domain D.

Examples. 1. It is easy to verify that three functions of four variables

$$u_1 = x_1^2 + x_2^2 + x_3^2 + x_4^2,$$
$$u_2 = x_1 + x_2 + x_3 + x_4,$$
$$u_3 = 2x_1x_2 + 2x_1x_3 + 2x_1x_4 + 2x_2x_3 + 2x_2x_4 + 2x_3x_4$$

are dependent in any domain D of a four-dimensional space because

$$u_1 = u_2^2 - u_3$$

for all the points (x_1, x_2, x_3, x_4) of that domain.

* In a particular case, we can take as D a certain neighbourhood of a fixed point M_0 of an n-dimensional space.

2. We shall now show that two functions of two variables $u_1 = x + y$ and $u_2 = x - y$ are independent in any domain D of the plane x, y containing the origin. It is clear that the function u_1 retains the constant value zero on the straight line $x + y = 0$ passing through the origin (Fig. 15.5). But the function u_2 has a variable value $u_2 = 2x$ on that line. Therefore, on the portion of the line lying within D, u_2 is obviously independent of u_1. It can be proved by complete analogy that on the portion of the line $x - y = 0$ lying within the domain D, $u_2 = 0$, $u_1 = 2x$ and, consequently, u_1 does not depend on u_2.

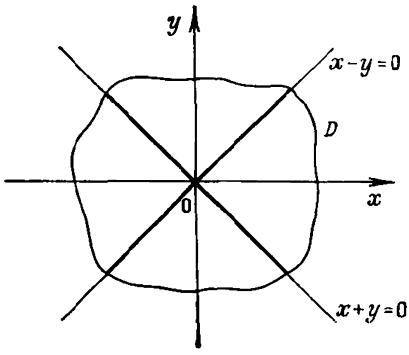


Fig. 15.5

Remark. The concept of a *linear dependence* of a function is usually introduced in the course of linear algebra: m functions $u_1, u_2, \ldots, u_m$ are said to be linearly dependent in the domain D if one of these functions can be expressed in the form of a linear function of the others for all the points of the domain D. It is clear that the linear dependence of functions is a particular case of the dependence of these functions, because if the functions $u_1, u_2, \ldots, u_m$ are linearly dependent in the domain D, they are dependent in that domain, but there are functions which are dependent in the domain D but are not linearly dependent in D (for instance, the functions written out in Example 1).

Theorem 15.3 (sufficient condition of independence of functions). *Assume that m functions of $n \geqslant m$ variables*

$$u_1 = \varphi_1(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n),$$
$$u_m = \varphi_m(x_1, \ldots x_m, x_{m+1}, \ldots, x_n)$$

are defined and differentiable in the neighbourhood of point $M_0(\overset{\circ}{x}_1, \ldots \overset{\circ}{x}_m, \overset{\circ}{x}_{m+1}, \ldots, \overset{\circ}{x}_n)$. Then, if the Jacobian of these functions in some m variables is nonzero at point M_0, the functions are independent in some neighbourhood of point M_0.

Proof. Without violating generality, we shall consider the Jacobian

$$\frac{D(u_1, u_2, \ldots, u_m)}{D(x_1, x_2, \ldots, x_m)} \tag{15.30}$$

to be nonzero at point M_0. We shall prove the theorem by contradiction. We shall suppose that the functions $u_1, u_2, \ldots, u_m$ are dependent in some neighbourhood of point M_0, that is, that for all the points of this neighbourhood one of these functions, say u_k, can be expressed in the form

$$u_k = \Phi(u_1, \ldots, u_{k-1}, u_{k+1}, \ldots, u_m),$$

where Φ is a certain differentiable function. Using the rule of differentiation of a composite function, we shall calculate the derivative of the function u_k in any one of the variables x_l ($l = 1, 2, \ldots, m$). We shall have

$$\frac{\partial u_k}{\partial x_l} = \frac{\partial \Phi}{\partial u_1}\frac{\partial u_1}{\partial x_l} + \ldots + \frac{\partial \Phi}{\partial u_{k-1}}\frac{\partial u_{k-1}}{\partial x_l} + \\ + \frac{\partial \Phi}{\partial u_{k+1}}\frac{\partial u_{k+1}}{\partial x_l} + \ldots + \frac{\partial \Phi}{\partial u_m}\frac{\partial u_m}{\partial x_l} \quad (l = 1, 2, \ldots, m). \tag{15.31}$$

If formulas (15.31) are taken for every value $l = 1, 2, \ldots, m$ at point M_0, they indicate that the kth row of Jacobian (15.30) is a linear combination of the remaining rows in coefficients equal, respectively, to $\frac{\partial \Phi}{\partial u_1}, \frac{\partial \Phi}{\partial u_2}, \ldots, \frac{\partial \Phi}{\partial u_{k-1}}, \frac{\partial \Phi}{\partial u_{k+1}}, \ldots, \frac{\partial \Phi}{\partial u_m}$. But in that case, Jacobian (15.30) is equal to zero at point M_0, which contradicts the hypothesis.

Example. The two functions $u_1 = x + y$ and $u_2 = x - y$ discussed above are independent in the neighbourhood of any point $M(x, y)$, since the Jacobian $\frac{D(u_1, u_2)}{D(x, y)} = \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} = -2 \neq 0$ everywhere.

15.4.2. Functional matrices and their applications. We shall again consider m functions of n variables (15.28):

$$\left.\begin{aligned} u_1 &= \varphi_1(x_1, x_2, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots \\ u_m &= \varphi_m(x_1, x_2, \ldots, x_n). \end{aligned}\right\} \tag{15.28}$$

We shall suppose this time that functions (15.28) *are defined and differentiable in some neighbourhood of point* $M_0(\overset{\circ}{x}_1, \ldots, \overset{\circ}{x}_n)$, *all the first-order partial derivatives of these functions being continuous at point* M_0 *itself.*

Let us compile from the partial derivatives of functions (15.28) the following functional matrix:

$$\left\| \begin{matrix} \frac{\partial \varphi_1}{\partial x_1} & \frac{\partial \varphi_1}{\partial x_2} & \ldots & \frac{\partial \varphi_1}{\partial x_n} \\ \frac{\partial \varphi_2}{\partial x_1} & \frac{\partial \varphi_2}{\partial x_2} & \ldots & \frac{\partial \varphi_2}{\partial x_n} \\ \ldots & \ldots & \ldots & \ldots \\ \frac{\partial \varphi_m}{\partial x_1} & \frac{\partial \varphi_m}{\partial x_2} & \ldots & \frac{\partial \varphi_m}{\partial x_n} \end{matrix} \right\| \tag{15.32}$$

containing m rows and n columns.

The following statement is true.

Theorem 15.4. *Let us consider the functional matrix* (15.32) *and assume that* (1) *its certain rth order minor* is nonzero at point* $M_0(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_n)$, (2) *all its minors of the* $(r+1)$*th order are equal to zero in some neighbourhood of point* M_0^{**}. *Then r functions represented in the rth order minor are independent in the neighbourhood of point* M_0, *and each of the other functions depends in this neighbourhood on the indicated r functions.*

Proof. Without loss of generality, we shall assume that different from zero at point M_0 is the minor positioned at the upper left-hand corner of matrix (15.32), i.e. distinct from zero is the determinant

$$\begin{vmatrix} \frac{\partial \varphi_1}{\partial x_1} & \cdots & \frac{\partial \varphi_1}{\partial x_r} \\ \cdots & \cdots & \cdots \\ \frac{\partial \varphi_r}{\partial x_1} & \cdots & \frac{\partial \varphi_r}{\partial x_r} \end{vmatrix}. \tag{15.33}$$

Then, the independence of the functions $u_1, u_2, \ldots, u_r$ in the neighbourhood of point M_0 follows immediately from Theorem 15.3. It remains to prove that any one of the functions $u_{r+1}, \ldots, u_m$*** depends on $u_1, u_2, \ldots, u_r$ in the neighbourhood of M_0. We shall prove, for instance, that u_{r+1} *depends on* $u_1, u_2, \ldots, u_r$ *in the neighbourhood of point* M_0. Let us concentrate on the first r functions (15.28). If we denote numbers of the form $\overset{\circ}{u}_1 = \varphi_1(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_n), \ldots, \overset{\circ}{u}_r = \varphi_r(\overset{\circ}{x}_1, \overset{\circ}{x}_2, \ldots, \overset{\circ}{x}_n)$ by $\overset{\circ}{u}_1, \ldots, \overset{\circ}{u}_r$, then everywhere in some neighbourhood of point $N_0(\overset{\circ}{u}_1, \ldots, \overset{\circ}{u}_r, \overset{\circ}{x}_1, \ldots, \overset{\circ}{x}_n)$ of the $(n+r)$-dimensional space the first r functions (15.28) are the unique and differentiable solution of the following system of equations****:

$$\left.\begin{aligned} F_1(u_1, \ldots, u_r, x_1, \ldots, x_n) &\equiv \varphi_1(x_1, \ldots, x_n) - u_1 = 0, \\ \cdots\cdots\cdots\

coinciding with minor (15.33), is nonzero at point N_0, system (15.34) can be *solved unambiguously with respect to* $x_1, \ldots, x_r$ in the neighbourhood of that point. To put it otherwise, everywhere in a sufficiently small neighbourhood of point N_0, system (15.34) possesses a unique and differentiable solution

$$\left.\begin{aligned} x_1 &= \psi_1(u_1, \ldots, u_r, x_{r+1}, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ x_r &= \psi_r(u_1, \ldots, u_r, x_{r+1}, \ldots, x_n). \end{aligned}\right\} \tag{15.35}$$

It should be emphasized that equalities (15.35) and the first r equalities (15.28) are completely equivalent in the neighbourhood of point N_0. In particular, if $x_1, x_2, \ldots, x_r$ specified by equations (15.35) are substituted into the first r equalities (15.28), these equalities turn into identities with respect to $x_{r+1}, \ldots, x_n, u_1, \ldots, u_r$. Differentiating these identities with respect to the variable x_l ($l = r+1, \ldots, n$) and taking into consideration that $u_1, \ldots, u_r$ do not depend on $x_{r+1}, \ldots, x_n$ we shall have

$$\frac{\partial\varphi_1}{\partial x_1}\frac{\partial\psi_1}{\partial x_l} + \ldots + \frac{\partial\varphi_1}{\partial x_r}\frac{\partial\psi_r}{\partial x_l} + \frac{\partial\varphi_1}{\partial x_l} = 0, \tag{15.36^1}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\frac{\partial\varphi_r}{\partial x_1}\frac{\partial\psi_1}{\partial x_l} + \ldots + \frac{\partial\varphi_r}{\partial x_r}\frac{\partial\psi_r}{\partial x_l} + \frac{\partial\varphi_r}{\partial x_l} = 0. \tag{15.36^r}$$

Note that equalities (15.36^1)-(15.36^r) are valid for all the values of the variables $x_1, \ldots, x_r, x_{r+1}, \ldots, x_n$ belonging to some neighbourhood of point M_0.

To make sure that the function u_{r+1} depends on $u_1, \ldots, u_r$ in some neighbourhood of point M_0, let us substitute the values of $x_1, \ldots, x_r$ specified by equations (15.35) into the $(r+1)$th equality (15.28). This will turn u_{r+1} into a function of the arguments $u_1, \ldots, u_r, x_{r+1}, \ldots, x_n$ since $u_{r+1} = \varphi_{r+1}(x_1, \ldots, x_r, x_{r+1}, \ldots, x_n) = \varphi_{r+1}[\psi_1(u_1, \ldots, u_r, x_{r+1}, \ldots, x_n), \ldots, \psi_r(u_1, \ldots, u_r, x_{r+1}, \ldots, x_n), x_{r+1}, \ldots, x_n] = \Phi(u_1, \ldots, u_r, x_{r+1}, \ldots, x_n)$ (this function will be denoted by Φ). It remains to prove that for all the values of the variables $x_1, \ldots, x_r, x_{r+1}, \ldots, x_n$ lying in a sufficiently small neighbourhood of point M_0 the function Φ *does not depend on* $x_{r+1}, \ldots, x_n$. For that purpose it is sufficient to prove that the equalities

$$\frac{\partial\Phi}{\partial x_l} = 0 \quad (l = r+1, \ldots, n) \tag{15.37}$$

hold true for all $x_1, \ldots, x_n$ from a sufficiently small neighbourhood of point M_0. Differentiating the function Φ with respect to the variable x_l ($l = r+1, \ldots, n$) as a composite function, we obtain

$$\frac{\partial\varphi_{r+1}}{\partial x_1}\frac{\partial\psi_1}{\partial x_l} + \ldots + \frac{\partial\varphi_{r+1}}{\partial x_r}\frac{\partial\psi_r}{\partial x_l} + \frac{\partial\varphi_{r+1}}{\partial x_l} = \frac{\partial\Phi}{\partial x_l}. \tag{15.36^{r+1}}$$

Let us now consider the following $(r+1)$th order minor of matrix (15.32):

$$\Delta = \begin{vmatrix} \frac{\partial \varphi_1}{\partial x_1} & \dots & \frac{\partial \varphi_1}{\partial x_r} & \frac{\partial \varphi_1}{\partial x_l} \\ \dots & \dots & \dots & \dots \\ \frac{\partial \varphi_r}{\partial x_1} & \dots & \frac{\partial \varphi_r}{\partial x_r} & \frac{\partial \varphi_r}{\partial x_l} \\ \frac{\partial \varphi_{r+1}}{\partial x_1} & \dots & \frac{\partial \varphi_{r+1}}{\partial x_r} & \frac{\partial \varphi_{r+1}}{\partial x_l} \end{vmatrix}. \tag{15.38}$$

By the hypothesis, this minor *is equal to zero everywhere in the neighbourhood of point* M_0. We shall multiply equalities (15.36^1-15.36^{r+1}) by the respective algebraic cofactors $\Delta_1, \dots, \Delta_r, \Delta_{r+1}$ of the elements of the last column of minor (15.38) and then add up all the equalities that we shall obtain. In accordance with the theorem stating that the sum of products of the elements of a given column by the respective algebraic cofactors of the elements of this (other) column is equal to the determinant (zero), we obtain*

$$\Delta = \frac{\partial \Phi}{\partial x_l} \Delta_{r+1}. \tag{15.39}$$

The symbol Δ in equality (15.39) denotes minor (15.38) equal to zero everywhere in the neighbourhood of point M_0, and the algebraic cofactor Δ_{r+1} coincides with minor (15.33) differing from zero at point M_0 and, hence, in some neighbourhood of that point**. We infer from equality (15.39) that equalities (15.37) are true everywhere in some neighbourhood of point M_0. We have proved the theorem.

Example. Let us return to the discussion of the dependence of the functions

$$u_1 = x_1^2 + x_2^2 + x_3^2 + x_4^2,$$
$$u_2 = x_1 + x_2 + x_3 + x_4,$$
$$u_3 = 2x_1x_2 + 2x_1x_3 + 2x_1x_4 + 2x_2x_3 + 2x_2x_4 + 2x_3x_4.$$

Functional matrix (15.32) has the form

$$\begin{Vmatrix} 2x_1 & 2x_2 & 2x_3 & 2x_4 \\ 1 & 1 & 1 & 1 \\ 2(x_2+x_3+x_4) & 2(x_1+x_3+x_4) & 2(x_1+x_2+x_4) & 2(x_1+x_2+x_3) \end{Vmatrix}.$$

* We repeat here the arguments carried out on pp. 573 – 574 in detail.

** Since all the partial derivatives appearing in minor (15.33) are continuous at point M_0, minor (15.33) itself is continuous at point M_0. But then, in accordance with the theorem on the stability of sign of a continuous function, this minor is nonzero, not only at point M_0 itself, but also in its certain neighbourhood.

It is easy to verify that all the third-order determinants are identically equal to zero, and that at any point of the space (x_1, x_2, x_3, x_4) whose all four coordinates x_1, x_2, x_3, x_4 coincide, at least one of the second-order determinants

$$\begin{vmatrix} 2x_1 & 2x_2 \\ 1 & 1 \end{vmatrix}, \quad \begin{vmatrix} 2x_1 & 2x_3 \\ 1 & 1 \end{vmatrix}, \quad \begin{vmatrix} 2x_1 & 2x_4 \\ 1 & 1 \end{vmatrix}$$

is nonzero. Consequently u_1 and u_2 are independent while u_3 depends on u_1, and u_2 in the neighbourhood of any one of the indicated points.

15.5. CONDITIONAL EXTREMUM

15.5.1. The concept of the conditional extremum. In 14.6 we were occupied with seeking the local extrema of a function whose arguments are not under any additional conditions. However, in mathematics and its applications we often come across a problem *of seeking the extrema of a function whose arguments satisfy the additional conditions of constraint.* Extrema of this kind are called *conditional* extrema to distinguish them from (unconditional) extrema we studied in 14.6.

Here is an example of a problem of seeking a conditional extremum. Let it be required to find the extremum of the function $u = x^2 + y^2$ under the condition that the arguments of this function comply with the constraint condition $x + y - 1 = 0$. Thus, the extrema of the function $u = x^2 + y^2$ are sought, *not on the entire plane Oxy, but only on the straight line* $x + y - 1 = 0$. To solve the problem we have posed, we shall substitute into the equation of the function $u = x^2 + y^2$ the value of y determined from the constraint equation $x + y - 1 = 0$. We shall thus reduce the problem we have to solve to that of seeking an unconditional extremum of the function $u = 2x^2 - 2x + 1$.

This unconditional extremum is easy to find: since $u' = 4\left(x - \frac{1}{2}\right)$, $u^{(2)} = 4$, the function $u = 2x^2 - 2x + 1$ possesses a minimum $u = \frac{1}{2}$ at $x = \frac{1}{2}$. Thus we see that the function $u = x^2 + y^2$ with constraint condition $x + y - 1 = 0$ possesses a conditional minimum $u = \frac{1}{2}$ at point $\left(\frac{1}{2}, \frac{1}{2}\right)$. Note that the unconditional minimum of the function $u = x^2 + y^2$ is attained at point (0, 0) and is equal to $u = 0$. However, as can be seen in Fig. 15.6, the minimum of the function $u = x^2 + y^2$ (whose graph is a paraboloid of revolution) on the entire Oxy plane does not coincide with its minimum on the straight line $x + y - 1 = 0$.

It is now time to pose a general problem of seeking a conditional extremum. Let it be required to find the extremum of the function

of $m + n$ variables

$$u = f(x_1, \ldots, x_n, y_1, \ldots, y_m) \tag{15.40}$$

given m constraint equations

$$\left.\begin{aligned} &F_1(x_1, \ldots, x_n, y_1, \ldots, y_m) = 0, \\ &F_2(x_1, \ldots, x_n, y_1, \ldots, y_m) = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &F_m(x_1, \ldots, x_n, y_1, \ldots, y_m) = 0. \end{aligned}\right\} \tag{15.41}$$

Let us elucidate, first of all, the notion of the conditional extremum of function (15.40) when we have constraints (15.41). We shall say that *under constraint conditions* (15.41), *function* (15.40) *possesses a conditional maximum (minimum) at point* $M_0(\overset{\circ}{x}_1, \ldots, \overset{\circ}{x}_n, \overset{\circ}{y}_1, \ldots, \overset{\circ}{y}_m)$, *whose coordinates satisfy constraint conditions* (15.41), *if there is a neighbourhood of point* M_0 *throughout which the value of function* (15.40) *at point* M_0 *is the greatest (the least) among its values at all the points whose coordinates satisfy conditions of constraint* (15.41).

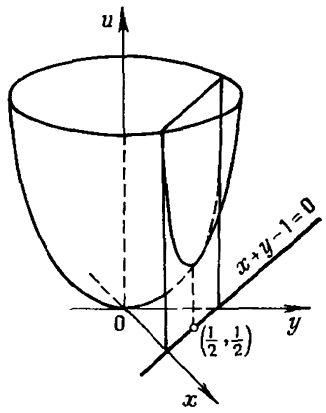


Fig. 15.6

To find the conditional extremum of function (15.40), under constraint conditions (15.41), let us assume that the functions on the left-hand sides of (15.41) are differentiable in some neighbourhood of point M_0 in question, the partial derivatives of the indicated functions with respect to $y_1, \ldots, y_m$ being continuous at point M_0 itself and the Jacobian

$$\frac{D(F_1, \ldots, F_m)}{D(y_1, \ldots, y_m)} \tag{15.42}$$

being nonzero.

In such a case, by virtue of Theorem 15.2, for sufficiently small positive numbers $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$, there is a neighbourhood of point $M_0'(\overset{\circ}{x}_1, \ldots, \overset{\circ}{x}_n)$ of the space of the variables $(x_1, \ldots, x_n)$ such that everywhere within this neighbourhood m functions

$$\left.\begin{aligned} y_1 &= \varphi_1(x_1, \ldots, x_n), \\ y_2 &= \varphi_2(x_1, \ldots, x_n), \\ &\ldots\ldots\ldots\ldots \\ y_m &= \varphi_m(x_1, \ldots, x_n) \end{aligned}\right\} \tag{15.43}$$

are defined which satisfy the conditions $|y_1 - \overset{\circ}{y}_1| < \varepsilon_1, \ldots, |y_m - \overset{\circ}{y}_m| < \varepsilon_m$ and which, under these conditions, are the

unique and differentiable solution of system of equations (15.41). Substituting functions (15.43) we have found into (15.40), we reduce the problem concerning the existence of a conditional extremum of function (15.40) at point M_0 under constraint conditions (15.41) to the problem of the existence of an unconditional extremum of the composite function

$$u = f[x_1, \ldots, x_n, \varphi_1(x_1, \ldots, x_n), \ldots, \varphi_m(x_1, \ldots, x_n)] = \\ = \Phi(x_1, \ldots, x_n) \tag{15.44}$$

of the arguments $x_1, \ldots, x_n$ at point M_0. The problem of the existence of an unconditional extremum of function (15.44) can be solved by the methods described in 14.6*. We have used a particular example to realize the general scheme of reducing a conditional extremum to an unconditional one. In what follows, we shall try, *without solving system* (15.41), to establish at least the necessary conditions of the existence of a conditional extremum at point M_0. Thus, assume that function (15.40) is differentiable at point M_0 and possesses a conditional extremum at that point when constraints (15.41) are given, or, which is the same, that function (15.44) has an unconditional extremum at point M_0'. In accordance with the facts established in 14.6, the necessary condition of the existence of an unconditional extremum of the function $u = \Phi(x_1, \ldots, x_n)$ at point M_0' is the equality to zero, at that point, of the differential of that function

$$du = \frac{\partial \Phi}{\partial x_1} dx_1 + \ldots + \frac{\partial \Phi}{\partial x_n} dx_n = 0, \tag{15.45}$$

the equality being identical with respect to $dx_1, \ldots, dx_n$. By virtue of the invariance of the form of the first differential and equality (15.44), formula (15.45) can be rewritten as

$$du = \frac{\partial f}{\partial x_1} dx_1 + \ldots + \frac{\partial f}{\partial x_n} dx_n + \frac{\partial f}{\partial y_1} dy_1 + \ldots + \frac{\partial f}{\partial y_m} dy_m = 0. \tag{15.46}$$

(In this formula all the partial derivatives are taken at point M_0.) It should be emphasized, however, that $dy_1, \ldots, dy_m$ in equality (15.46) are the differentials of functions (15.43) so that equality (15.46) is not an identity with respect to $dy_1, \ldots, dy_m$. Let us assume that substituted into constraint equations (15.41) are functions (15.43) which are the solution of system (15.41). In that case, equations (15.41)

* To do that, certain conditions must, of course, be imposed on function (15.40).

turn into identities and we obtain, by differentiating these identities,

$$\left.\begin{aligned} &\frac{\partial F_1}{\partial x_1}dx_1+\ldots+\frac{\partial F_1}{\partial x_n}dx_n+\frac{\partial F_1}{\partial y_1}dy_1+\ldots+\frac{\partial F_1}{\partial y_m}dy_m=0,\\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ &\frac{\partial F_m}{\partial x_1}dx_1+\ldots+\frac{\partial F_m}{\partial x_n}dx_n+\frac{\partial F_m}{\partial y_1}dy_1+\ldots+\frac{\partial F_m}{\partial y_m}dy_m=0. \end{aligned}\right\} \tag{15.47}$$

Since, according to the assumption, Jacobian (15.42) is different from zero at point M_0, the quantities $dy_1, \ldots, dy_m$ from the linear system (15.47) can be expressed as the linear functions $dx_1, \ldots, dx_n$. If we find these expressions and substitute them into (15.46), then, combining the terms containing $dx_1, \ldots, dx_n$ in the resulting equality, we shall have

$$A_1dx_1+\ldots+A_ndx_n=0, \tag{15.48}$$

where $A_1, \ldots, A_n$ denote some rational functions of the partial derivatives $f, F_1, \ldots, F_m$ at point M_0. Since equality (15.48) contains *only the differentials of independent variables*, we infer from this equality that $A_1=0, \ldots, A_n=0$. Adding the m constraint equations (15.41) to these equalities, we shall obtain the *necessary conditions* of the existence of the conditional extremum of function (15.40) under constraint conditions (15.41), in the form

$$A_1=0, \ldots, A_n=0, \quad F_1=0, \ldots, F_m=0. \tag{15.49}$$

Equalities (15.49) are a system of $m+n$ equations for the determination of $m+n$ coordinates of the point of a possible extremum.

15.5.2. Lagrange's method of multipliers. When exposing the method of seeking points of a possible conditional extremum, we have violated the symmetry as concerns the variables $x_1, \ldots, x_n$, $y_1, \ldots, y_m$. Some of these variables $x_1, \ldots, x_n$, were regarded as independent, the others as functions of these variables. In some cases this complicates the calculations. The method proposed by Lagrange serves to symmetrize the role played by variables. We devote this subsection to the exposition of Lagrange's method. Let us multiply equalities (15.47) by the arbitrary (as yet undetermined) constant factors $\lambda_1, \ldots, \lambda_m$ respectively. Then we shall add termwise the equalities obtained to equality (15.46). The resulting equality will be of the form

$$\frac{\partial \Psi}{\partial x_1}dx_1+\ldots+\frac{\partial \Psi}{\partial x_n}dx_n+\frac{\partial \Psi}{\partial y_1}dy_1+\ldots+\frac{\partial \Psi}{\partial y_m}dy_m=0 \tag{15.50}$$

where the symbol $\Psi(x_1, \ldots, x_n, y_1, \ldots, y_m)$ denotes the following function:

$$\Psi=f+\lambda_1F_1+\ldots+\lambda_mF_m. \tag{15.51}$$

In what follows, this function will be called a *Lagrangian function.* Assuming that the conditions formulated in the previous subsection are satisfied for functions (15.41) and that function (15.40) is differentiable, we shall choose the factors $\lambda_1, \ldots, \lambda_m$ such that the equalities

$$\frac{\partial \Psi}{\partial y_1} = 0, \quad \ldots \quad \frac{\partial \Psi}{\partial y_m} = 0 \tag{15.52}$$

are satisfied. This can obviously be done since equalities (15.52) lead to the linear system

$$\left.\begin{aligned} &\frac{\partial f}{\partial y_1} + \lambda_1 \frac{\partial F_1}{\partial y_1} + \ldots + \lambda_m \frac{\partial F_m}{\partial y_1} = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\frac{\partial f}{\partial y_m} + \lambda_1 \frac{\partial F_1}{\partial y_m} + \ldots + \lambda_m \frac{\partial F_m}{\partial y_m} = 0, \end{aligned}\right\}$$

whose determinant (Jacobian (15.42)) is nonzero. By virtue of equalities (15.52), equality (15.50) takes the form

$$\frac{\partial \Psi}{\partial x_1} dx_1 + \ldots + \frac{\partial \Psi}{\partial x_n} dx_n = 0. \tag{15.53}$$

Since under the assumptions made above, the variables $x_1, \ldots, x_n$ are *independent*, we infer from equality (15.53) that

$$\frac{\partial \Psi}{\partial x_1} = 0, \ldots \frac{\partial \Psi}{\partial x_n} = 0. \tag{15.54}$$

Adding the constraint equations (15.41) to equations (15.52) and (15.54), we obtain a system of $n + 2m$ equations

$$\left.\begin{aligned} &\frac{\partial \Psi}{\partial x_1} = 0, \ \ldots, \frac{\partial \Psi}{\partial x_n} = 0, \ \frac{\partial \Psi}{\partial y_1} = 0, \ \ldots, \frac{\partial \Psi}{\partial y_m} = 0, \\ &F_1 = 0, \ \ldots, \ F_m = 0 \end{aligned}\right\} \tag{15.55}$$

for the determination of $n + m$ coordinates of points of a possible conditional extremum and m multipliers $\lambda_1, \ldots, \lambda_m$. In practice, when this method is realized, the proceedure is the following. The Lagrangian function (15.51) is compiled and then *points of a possible unconditional extremum are found* for that function. To eliminate the multipliers $\lambda_1, \ldots, \lambda_m$, constraint equations (15.41) are made use of. This technique of seeking points of a possible conditional extremum is legitimate since it leads precisely to a system of $n + 2M$ equations (15.55). We shall proceed with discussing Lagrange's method of multipliers in 15.5.4.

15.5.3. Sufficient conditions. We shall consider here one of the ways of additional investigation of points of a possible conditional extremum. Assume that the necessary conditions of extremum (15.55) are satisfied at point M_0. We shall additionally require that func-

tions (15.40) and (15.41) be twice differentiable in the neighbourhood of point M_0 and that all the second-order partial derivatives be continuous at point M_0 itself. The construction of the Lagrangian function (15.51) implies that *under constraint conditions* (15.41), the extrema of function (15.40) and of the Lagrangian function coincide*. But then it follows from the results obtained in 14.6 that to attain the sufficient condition of the extremum of function (15.40) at point M_0, under constraint conditions (15.41), *conditions* (15.55) *must be supplemented with the requirement that* $d^2\Psi$ *should have a definite sign at that point.* In that case, in accordance with the results of 14.6, we can verify the existence of a minimum at point M_0 if $d^2\Psi\,|_{M_0} > 0$ under the conditions of constraint (15.41) and of maximum if $d^2\Psi\,|_{M_0} < 0$. Some practical remarks are due now. Note, first of all, that *the second differential* $d^2\Psi$ *can be calculated at the given point* M_0 *of a possible extremum under the assumption that all the variables* $x_1, \ldots, x_n, y_1, \ldots, y_m$ *are independent.* Indeed, in a general case, the second differential $d^2\Psi$ of the function Ψ does not possess the invariance of the form and so, were the dependence of $y_1, \ldots, y_m$ on $x_1, \ldots, x_n$ taken into account, it should have been determined by the equality

$$d^2\Psi = \left(dx_1 \frac{\partial}{\partial x_1} + \ldots + dx_n \frac{\partial}{\partial x_n} + dy_1 \frac{\partial}{\partial y_1} + \ldots + dy_m \frac{\partial}{\partial y_m}\right)^2 \Psi + \\ + \frac{\partial \Psi}{\partial y_1} d^2 y_1 + \ldots + \frac{\partial \Psi}{\partial y_m} d^2 y_m.$$

But at the point M_0 of a possible extremum there hold equalities

$$\frac{\partial \Psi}{\partial y_1} = 0, \ \ldots, \ \frac{\partial \Psi}{\partial y_m} = 0,$$

so that $d^2\Psi$ is determined by the same formula

$$d^2\Psi = \left(dx_1 \frac{\partial}{\partial x_1} + \ldots + dx_n \frac{\partial}{\partial x_n} + dy_1 \frac{\partial}{\partial y_1} + \ldots + dy_m \frac{\partial}{\partial y_m}\right)^2 \Psi, \tag{15.56}$$

that is used in the case where all the variables $x_1, \ldots, x_n, y_1, \ldots, y_m$ are independent. It should also be pointed out that since we need to establish whether $d^2\Psi$ is positive or negative only *under the conditions of constraint* (15.41), it follows that to carry out the calculations we must replace $dy_1, \ldots, dy_m$ in formula (15.56) for $d^2\Psi$ by their values determined from system (15.47). The next step is to decide whether $d^2\Psi$ is positive or negative at the given point M_0. Now we can consider an example.

* This follows from the fact that under constraint conditions (15.41) the difference $f(M) - f(M_0)$ coincides with the difference $\Psi(M) - \Psi(M_0)$.

15.5.4. Example. Suppose we must find the maximum and the minimum value of the determinant

$$\Delta = \begin{vmatrix} x_1 & y_1 & \dots & z_1 \\ x_2 & y_2 & \dots & z_2 \\ \dots & \dots & \dots & \dots \\ x_n & y_n & \dots & z_n \end{vmatrix}, \tag{15.57}$$

with the sum of the squares of the elements of each row of the determinant being known. The problem reduces to seeking the extremal values of the function of n^2 variables (15.57) under the following n conditions of constraint*

$$\left.\begin{aligned} x_1^2 + y_1^2 + \dots + z_1^2 &= h_1, \\ x_2^2 + y_2^2 + \dots + z_2^2 &= h_2, \\ \dots\dots\dots\dots\dots & \\ x_n^2 + y_n^2 + \dots + z_n^2 &= h_n, \end{aligned}\right\} \tag{15.58}$$

where $h_1, h_2, \dots, h_n$ are given positive numbers**. To solve the problem, let us discuss and solve a simpler problem. *Let us fix the elements of all the rows of determinant* (15.57) *at the extremal point, with the only exception of the kth row.* In that case, determinant (15.57) can be regarded as a function of n variables $x_k, y_k, \dots, z_k$. An explicit expression for this function can be obtained by expanding the determinant by the elements of the kth row:

$$\Delta = x_k X_k + y_k Y_k + \dots + z_k Z_k. \tag{15.59}$$

Here $X_k, Y_k, \dots, Z_k$ denote the algebraic cofactors of the corresponding elements $x_k, y_k, \dots, z_k$. Since the elements of all the rows of the determinant, except for the kth row, are fixed at the extremal point, $X_k, Y_k, \dots, Z_k$ can be regarded as constant numbers. Let us pose the problem of seeking the *extrema of function* (15.59) *under one condition of constraint****

$$x_k^2 + y_k^2 + \dots + z_k^2 = h_k. \tag{15.60}$$

To solve this problem, we compile a Lagrangian function

$$\Psi = x_k X_k + y_k Y_k + \dots + z_k Z_k + \lambda (x_k^2 + y_k^2 + \dots + z_k^2 - h_k), \tag{15.61}$$

---

* The *existence* of the extremal values does not arouse any doubt, since function (15.57) is a *continuous* function of its n^2 variables on the closed set (15.58).

** We omit the trivial case where at least one of the numbers $h_1, h_2, \dots, h_n$ is equal to zero. Then determinant (15.57) is identically equal to zero.

*** We take the kth condition (15.58).

and decide the question as to the existence of an unconditional extremum for that function. From the conditions

$$\frac{\partial \Psi}{\partial x_k} = X_k + 2\lambda x_k = 0,$$

$$\frac{\partial \Psi}{\partial y_k} = Y_k + 2\lambda y_k = 0, \ \ldots, \ \frac{\partial \Psi}{\partial z_k} = Z_k + 2\lambda z_k = 0$$

we find the coordinates of the point of a possible extremum

$$x_k = -\frac{X_k}{2\lambda}, \ y_k = -\frac{Y_k}{2\lambda}, \ \ldots, \ z_k = -\frac{Z_k}{2\lambda}. \tag{15.62}$$

The constant factor λ can be easily eliminated from constraint (15.60). The constraint yields two values:

$$\lambda_1 = \sqrt{\frac{X_k^2 + Y_k^2 + \ldots + Z_k^2}{4h_k}} > 0, \ \lambda_2 = -\sqrt{\frac{X_k^2 + Y_k^2 + \ldots + Z_k^2}{4h_k}} < 0.$$

(Here again we omit the trivial case of $X_k, Y_k, \ldots, Z_k$ being equal to zero, since in that case determinant (15.59) is identically equal to zero.)

We thus obtain two points of a possible extremum:

$$M_1\left(-\frac{X_k}{2\lambda_1}, \ -\frac{Y_k}{2\lambda_1}, \ \ldots, \ -\frac{Z_k}{2\lambda_1}\right) \text{ and}$$

$$M_2\left(-\frac{X_k}{2\lambda_2}, \ -\frac{Y_k}{2\lambda_2}, \ \ldots, \ -\frac{Z_k}{2\lambda_2}\right).$$

Let us prove that a conditional minimum is realized at point M_1 and a conditional maximum at point M_2. For that purpose, let us calculate the second differential $d^2\Psi$ of the Lagrangian function (15.61). It is easy to see that

$$d^2\Psi = 2\lambda\left[(dx_k)^2 + (dy_k)^2 + \ldots + (dz_k)^2\right].$$

It follows from the last formula that $d^2\Psi$ is a positive definite quadratic form at $\lambda = \lambda_1 > 0$ (i.e. at point M_1) and a negative definite form at $\lambda = \lambda_2 < 0$ (i.e. at point M_2). Thus, under constraint condition (15.60), function (15.59) possesses a conditional minimum at point M_1 and a conditional maximum at point M_2. Without calculating the greatest and the least value of function (15.59) under condition (15.60), we shall point out that at the point at which these values are attained equalities (15.62) hold true and, consequently, the equalities

$$\frac{x_k}{X_k} = \frac{y_k}{Y_k} = \ldots = \frac{z_k}{Z_k} \tag{15.63}$$

are valid. Let us now return to the calculation of the extremal values of determinant (15.57) under the condition of n constraints (15.58). Retaining the meaning of the accepted notation $X_k, Y_k, \ldots, Z_k$,

we can write

$$xX_{ik} + y_iY_k + \ldots + z_iZ_k = 0 \quad (i \neq k) \tag{15.64}$$

in accordance with the known property of a determinant for any i distinct from k. We find from (15.63) and (15.64) that

$$x_ix_k + y_iy_k + \ldots + z_iz_k = 0 \quad (\text{at } i \neq k). \tag{15.65}$$

Equalities (15.58) and (15.65) and the rule of multiplication of determinants allow us to infer that when the determinant Δ is multiplied by itself, a determinant results whose all elements are equal to zero with the exception of the elements of the principal diagonal which are equal, respectively, to $h_1, h_2, \ldots, h_n$. Thus we have

$$\Delta \cdot \Delta = \Delta^2 = h_1h_2 \ldots h_n.$$

Consequently, under conditions (15.58), the greatest and the least value of determinant (15.57) are equal to $+\sqrt{h_1h_2 \ldots h_n}$ and $-\sqrt{h_1h_2 \ldots h_n}$ respectively.
We thus obtain

$$|\Delta| \leqslant \sqrt{h_1h_2 \ldots h_n},$$

or

$$|\Delta| \leqslant \sqrt{(x_1^2 + \ldots + z_1^2)(x_2^2 + \ldots + z_2^2) \ldots (x_n^2 + \ldots + z_n^2)}. \tag{15.66}$$

The last inequality valid for the arbitrary determinant (15.57) is known as *Hadamard* inequality.*

Remark. At $n = 3$, Hadamard inequality (15.66) admits of a simple geometric interpretation. In this case $|\Delta|$ is the volume of a parallelepiped constructed on the line segments OA_1, OA_2 and OA_3 connecting the origin O with points $A_1(x_1, y_1, z_1)$, $A_2(x_2, y_2, z_2)$, $A_3(x_3, y_3, z_3)$. Inequality (15.65) states that of all the parallelepipeds with edges of a given length the rectangular parallelepiped has the greatest volume.

SUPPLEMENT TO CHAPTER 15

CHANGE OF VARIABLES

The problem of a change of variables can be encountered in some sections of the analysis and in other branches of mathematics. The problem consists in the following. Suppose we are given an expression

$$F\left(x, y, z, \frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}, \ldots\right), \tag{15.67}$$

* Jacques Hadamard (1865-1963), a French mathematician.

containing the independent variables x, y, the function $z = z(x, y)$ and its partial derivatives. Instead of the independent variables x and y and the function $z = z(x, y)$, new independent variables u and v and a new function $w = w(u, v)$ are introduced and the relationships are defined which express u, v and w in terms of x, y and z:

$$\left.\begin{aligned} u &= \varphi(x, y, z), \\ v &= \psi(x, y, z), \\ w &= \chi(x, y, z). \end{aligned}\right\} \tag{15.68}$$

We are to transform expression (15.67) to the new variables u, v and w. We shall assume that functions (15.68) are differentiable a sufficient number of times, and that system (15.68) can be solved with respect to x, y and z and the first two equations (15.68), with respect to x and y.

It is evident that to solve the problem, it is sufficient to express [the partial derivatives $\frac{\partial z}{\partial x}$, $\frac{\partial z}{\partial y}$, ... in terms of u, v, w, $\frac{\partial w}{\partial u}$, $\frac{\partial w}{\partial v}$, ... This is how it can be done.

Bearing in mind that $z = z(x, y)$ and $w = w(u, v)$, we shall write the first differentials of functions (15.68). We obtain

$$du = \frac{\partial \varphi}{\partial x}\,dx + \frac{\partial \varphi}{\partial y}\,dy + \frac{\partial \varphi}{\partial z}\,dz = \frac{\partial \varphi}{\partial x}\,dx + \frac{\partial \varphi}{\partial y}\,dy + \frac{\partial \varphi}{\partial z}\left(\frac{\partial z}{\partial x}\,dx + \frac{\partial z}{\partial y}\,dy\right), \tag{15.69}$$

$$dv = \frac{\partial \psi}{\partial x}\,dx + \frac{\partial \psi}{\partial y}\,dy + \frac{\partial \psi}{\partial z}\,dz = \frac{\partial \psi}{\partial x}\,dx + \frac{\partial \psi}{\partial y}\,dy + \frac{\partial \psi}{\partial z}\left(\frac{\partial z}{\partial x}\,dx + \frac{\partial z}{\partial y}\,dy\right), \tag{15.70}$$

$$dw = \frac{\partial w}{\partial u}\,du + \frac{\partial w}{\partial v}\,dv = \frac{\partial \chi}{\partial x}\,dx + \frac{\partial \chi}{\partial y}\,dy + \frac{\partial \chi}{\partial z}\left(\frac{\partial z}{\partial x}\,dx + \frac{\partial z}{\partial y}\,dy\right). \tag{15.71}$$

Substituting into (15.71) du and dv specified by formulas (15.69) and (15.70) and equating the coefficients in dx and dy, we obtain a system of two equations

$$\left.\begin{aligned} \frac{\partial w}{\partial u}\left[\frac{\partial \varphi}{\partial x} + \frac{\partial \varphi}{\partial z}\frac{\partial z}{\partial x}\right] + \frac{\partial w}{\partial v}\left[\frac{\partial \psi}{\partial x} + \frac{\partial \psi}{\partial z}\frac{\partial z}{\partial x}\right] &= \frac{\partial \chi}{\partial x} + \frac{\partial \chi}{\partial z}\frac{\partial z}{\partial x}, \\ \frac{\partial w}{\partial u}\left[\frac{\partial \varphi}{\partial y} + \frac{\partial \varphi}{\partial z}\frac{\partial z}{\partial y}\right] + \frac{\partial w}{\partial v}\left[\frac{\partial \psi}{\partial y} + \frac{\partial \psi}{\partial z}\frac{\partial z}{\partial y}\right] &= \frac{\partial \chi}{\partial y} + \frac{\partial \chi}{\partial z}\frac{\partial z}{\partial x} \end{aligned}\right\}$$

which can be used to express $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$:

$$\frac{\partial z}{\partial x} = -\frac{\frac{\partial w}{\partial u}\frac{\partial \varphi}{\partial x} + \frac{\partial w}{\partial v}\frac{\partial \psi}{\partial x} - \frac{\partial \chi}{\partial x}}{\frac{\partial w}{\partial u}\frac{\partial \varphi}{\partial z} + \frac{\partial w}{\partial v}\frac{\partial \psi}{\partial z} - \frac{\partial \chi}{\partial z}},$$

$$\frac{\partial z}{\partial y} = -\frac{\frac{\partial w}{\partial u}\frac{\partial \varphi}{\partial y} + \frac{\partial w}{\partial v}\frac{\partial \psi}{\partial y} - \frac{\partial \chi}{\partial y}}{\frac{\partial w}{\partial u}\frac{\partial \varphi}{\partial z} + \frac{\partial w}{\partial v}\frac{\partial \psi}{\partial z} - \frac{\partial \chi}{\partial z}}.$$

If expression (15.67) also depends on the second-order partial derivatives, then, to determine these derivatives in terms of the partial derivatives of w with respect to u and v, we must write the first differentials of the first-order derivatives

which we have calculated.

Remark 1. A similar change of variables can be carried out in the case where the old derivatives are connected with the new ones not by relations (15.68) but by implicit relations of the form*

$$\left.\begin{aligned}&\Phi_1(u, v, w, x, y, z)=0,\\&\Phi_2(u, v, w, x, y, z)=0,\\&\Phi_3(u, v, w, x, y, z)=0.\end{aligned}\right\} \tag{15.72}$$

Remark 2. We have restricted the argument to the case of two variables to abbreviate the writing. The indicated technique is applicable in the case of any number of independent variables (and, in particular, in the case of one independent variable).

Example. Assume that z is a function of the variables x and y. It is required to transform the expression**

$$F=\Delta z=\frac{\partial^2 z}{\partial x^2}+\frac{\partial^2 z}{\partial y^2}$$

to the polar coordinates u and v:

$$x=u\cos v, \quad y=u\sin v. \tag{15.73}$$

Pay attention to the fact that in this example we have carried out only a change of independent variables. The function z has remained unchanged. It follows from (15.73) that

$$dx=\cos v\,du-u\sin v\,dv, \quad dy=\sin v\,du+u\cos v\,dv,$$

therefore, from the relations

$$dz=\frac{\partial z}{\partial x}dx+\frac{\partial z}{\partial y}dy=\frac{\partial z}{\partial u}du+\frac{\partial z}{\partial v}dv$$

we have

$$\frac{\partial z}{\partial x}(\cos v\,du-u\sin v\,dv)+\frac{\partial z}{\partial y}(\sin v\,du+u\cos v\,dv)=\frac{\partial z}{\partial u}du+\frac{\partial z}{\partial v}dv.$$

Equating the coefficients in du and dv, we find

$$\frac{\partial z}{\partial x}\cos v+\frac{\partial z}{\partial y}\sin v+\frac{\partial z}{\partial u},$$

$$-\frac{\partial z}{\partial x}u\sin v+\frac{\partial z}{\partial y}u\cos v=\frac{\partial z}{\partial v}.$$

Hence

$$\left.\begin{aligned}\frac{\partial z}{\partial x}&=\cos v\frac{\partial z}{\partial u}-\frac{1}{u}\sin v\frac{\partial z}{\partial v},\\\frac{\partial z}{\partial y}&=\sin v\frac{\partial z}{\partial u}+\frac{1}{u}\cos v\frac{\partial z}{\partial v}.\end{aligned}\right\} \tag{15.74}$$

---

* Which admit, of course, of solvability both with respect to u, v and w and with respect to x, y and z.

** This expression is called the *Laplacian operator* (or simply the *Laplacian*). It plays an important role in mathematics and its applications. Pierre Simon Laplace (1749-1827), a French astronomer, mathematician and physicist.

Now we shall find $\frac{\partial^2 z}{\partial x^2}$. Since $\frac{\partial^2 z}{\partial x^2} = \frac{\partial\left(\frac{\partial z}{\partial x}\right)}{\partial x}$, we find from the first formula (15.74)

$$\frac{\partial^2 z}{\partial x^2} = \frac{\partial\left(\frac{\partial z}{\partial x}\right)}{\partial x} = \cos v \frac{\partial\left(\frac{\partial z}{\partial x}\right)}{\partial u} - \frac{1}{u}\sin v \frac{\partial\left(\frac{\partial z}{\partial x}\right)}{\partial v} =$$

$$= \cos v \frac{\partial}{\partial u}\left[\cos v \frac{\partial z}{\partial u} - \frac{1}{u}\sin v \frac{\partial z}{\partial v}\right] - \frac{1}{u}\sin v \frac{\partial}{\partial v}\left[\cos v \frac{\partial z}{\partial u} - \frac{1}{u}\sin v \frac{\partial z}{\partial v}\right].$$

The calculations result in

$$\frac{\partial^2 z}{\partial x^2} = \cos^2 v \frac{\partial^2 z}{\partial u^2} - \frac{2}{u}\sin v \cos v \frac{\partial^2 z}{\partial u\, \partial v} + \frac{1}{u^2}\sin^2 v \frac{\partial^2 z}{\partial v^2} +$$

$$+ \frac{1}{u}\sin^2 v \frac{\partial z}{\partial u} + \frac{2}{u^2}\sin v \cos v \frac{\partial z}{\partial v}.$$

By analogy, we find from the second formula (15.74)

$$\frac{\partial^2 z}{\partial y^2} = \frac{\partial\left(\frac{\partial z}{\partial y}\right)}{\partial y} = \sin v \frac{\partial\left(\frac{\partial z}{\partial y}\right)}{\partial u} + \frac{1}{u}\cos v \frac{\partial\left(\frac{\partial z}{\partial y}\right)}{\partial v} =$$

$$= \sin^2 v \frac{\partial^2 z}{\partial u^2} + \frac{2}{u}\sin v \cos v \frac{\partial^2 z}{\partial u\, \partial v} + \frac{1}{u^2}\cos^2 v \frac{\partial^2 z}{\partial v^2} +$$

$$+ \frac{1}{u}\cos^2 v \frac{\partial z}{\partial u} - \frac{2}{u^2}\sin v \cos v \frac{\partial z}{\partial v}.$$

Thus, in the polar coordinates u and v, the Laplacian $\Delta z = \frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2}$ has the following form:

$$\Delta z = \frac{\partial^2 z}{\partial u^2} + \frac{1}{u}\frac{\partial z}{\partial u} + \frac{1}{u^2}\frac{\partial^2 z}{\partial v^2}.$$

CHAPTER 16

SOME GEOMETRIC APPLICATIONS OF DIFFERENTIAL CALCULUS

16.1. THE ENVELOPE AND THE DISCRIMINANT CURVE OF A ONE-PARAMETER FAMILY OF PLANE CURVES

16.1.1. Preliminaries. We shall specify plane curves either by the parametric equations

$$x = \varphi(\alpha), \quad y = \psi(\alpha), \tag{16.1}$$

where α is a certain parameter, or by equations of the form

$$F(x, y) = 0. \tag{16.2}$$

It what follows we shall need the concepts of a *regular* (*nonsingular*) point and a *singular* point (singularity) of a curve*.

Suppose a curve L is specified by parametric equations (16.1) and the functions $x = \varphi(\alpha)$ and $y = \psi(\alpha)$ possess continuous derivatives at $\alpha = \alpha_0$. Point $M_0(x_0, y_0)$ of the curve L whose coordinates x_0 and y_0 are equal, respectively, to $\varphi(\alpha_0)$ and $\psi(\alpha_0)$ will be called a *regular* point if

$$\varphi'^2(\alpha_0) + \psi'^2(\alpha_0) \neq 0. \tag{16.3}$$

Now if the relation

$$\varphi'^2(\alpha_0) + \psi'^2(\alpha_0) = 0 \tag{16.4}$$

holds for $\alpha = \alpha_0$, then point M_0 will be called a *singular* point of the curve L.

Let the curve L be specified by equation (16.2), the function $F(x, y)$ being differentiable in some neighbourhood of point $M_0(x_0, y_0)$ of that curve and possessing continuous partial derivatives with respect to x and y at the indicated point. Point $M_0(x_0, y_0)$ is called a *regular* point of the curve L if the relation

$$F'^2_x + F'^2_y \neq 0 \tag{16.5}$$

holds at this point. Now if at point M_0 there holds a relation

$$F'^2_x + F'^2_y = 0, \tag{16.6}$$

then point M_0 is called a *singular* point of the curve L. Let us verify the fact that *if point M_0 of the curve L is regular, then, in some neighbourhood of this point, the curve L is either a graph of a certain differ-*

* The concepts of a regular point and a singular point for curves of the form (16.2) have been introduced in 15.2.3.

entiable function $y = f(x)$ *or a graph of a certain differentiable function* $x = g(y)$. Indeed, assume that the curve L is specified by parametric equations (16.1) and, in addition, condition (16.3) is fulfilled. It follows from the continuity of the derivatives $\varphi'(\alpha)$ and $\psi'(\alpha)$ at $\alpha = \alpha_0$ and from condition (16.3) that in some neighbourhood of α_0 at least one of these derivatives, say $\varphi'(\alpha)$, is nonzero. Then $x = \varphi(\alpha)$ is a differentiable strictly monotone function in the indicated neighbourhood. Under these conditions, there exists a differentiable monotone inverse function $\alpha = \varphi^{-1}(x)$. Substituting this function into the expression $y = \psi(\alpha)$, we shall verify that in a certain neighbourhood of point M_0 the curve L is a graph of the differentiable function $y = f(x) = \psi[\varphi^{-1}(x)]$. For the case where the curve is specified by equation (16.2), the validity of the assertion we have formulated follows from the fact that Theorem 15.1 on implicit functions works in the neighbourhood of a regular point and, therefore, the portion of the curve adjacent to a regular point is a graph of the differentiable function $y = f(x)$ or the function $x = g(y)$.

Remark 1. In geometry, point M_0 of the curve L is called regular if in some neighbourhood of that point the curve L is a graph of some differentiable function, and a singular point if in any neighbourhood of that point the curve L cannot be represented as a graph of a differentiable function. We have seen that a point of the curve L which is regular in accordance with our definition is also regular from the point of view of geometry. But we can cite some examples where a point of a curve being a singular point in accordance with our definition is a regular point from the point of view of geometry. So we see that our definition of a regular point is narrower than the geometric definition but is more convenient for applications.

We shall now introduce the notion of *tangency* of the curves L_1 and L_2 at their common point M_0. We shall say that the curves L_1 and L_2 are *tangent* at their common point M_0 if both curves possess tangent lines at point M_0 and the tangent lines coincide.

Later on we shall need the *condition of tangency* of two curves L_1 and L_2.

Assume that the curve L_1 is specified by equation (16.1) and the curve L_2 by equation (16.2), point $M_0(x_0, y_0)$ being their common point (in this case the coordinates x_0 and y_0 correspond to the value $\alpha = \alpha_0$ of the parameter α). We shall also assume that point M_0 is a *regular* point of the curves L_1 and L_2 and that the curves are *tangent* at point M_0. Then these curves are graphs of differentiable functions in the neighbourhood of M_0. For definiteness, we shall assume that L_1 and L_2 are graphs of the functions $y = f_1(x)$ and $y = f_2(x)$. Since by the hypothesis the curves L_1 and L_2 are tangent at point $M_0(x_0, y_0)$, the slopes of the tangents at M_0 to the graphs of the functions $f_1(x)$ and $f_2(x)$ are equal, i.e.

$$f_1'(x_0) = f_2'(x_0). \tag{16.7}$$

Using formulas for the differentiation of functions specified parametrically (see 5.11) and formulas for the differentiation of implicit functions (see 15.2.2), we obtain

$$f_1'(x_0)=\frac{\psi'(\alpha_0)}{\varphi'(\alpha_0)},\quad f_2'(x_0)=-\frac{F_x'(M_0)}{F_y'(M_0)}. \tag{16.8}$$

Formulas (16.8) allow equality (16.7) to take on the form

$$\frac{\psi'(\alpha_0)}{\varphi'(\alpha_0)}=-\frac{F_x'(M_0)}{F_y'(M_0)}, \tag{16.9}$$

or

$$F_x'(M_0)\,\varphi'(\alpha_0)+F_y'(M_0)\,\psi'(\alpha_0)=0. \tag{16.10}$$

In what follows, we shall call the last relation *the condition of tangency at point M_0 of the curves L_1 and L_2* specified by equations (16.1) and (16.2) respectively. Omitting the arguments of the functions and using the designations $\varphi'=\frac{dx}{d\alpha}$ and $\psi'=\frac{dy}{d\alpha}$, we shall write the condition of tangency in the following form:

$$F_x'\frac{dx}{d\alpha}+F_y'\frac{dy}{d\alpha}=0. \tag{16.11}$$

Remark 2. If the condition of tangency (16.10) (or (16.11), what is the same) is fulfilled at the common point M_0 of the curves L_1 and L_2 specified by equations (16.1) and (16.2) respectively, and if, in addition, point M_0 is a regular point of the curves L_1 and L_2, then the curves L_1 and L_2 are tangent at point M_0. Indeed, relations (16.3) and (16.5) and condition (16.10) yield either condition (16.9) or condition $\frac{\varphi'(\alpha_0)}{\psi'(\alpha_0)}=-\frac{F_y'(M_0)}{F_x'(M_0)}$, that is, the equality of the slopes of the tangents at the common point M_0 of the curves L_1 and L_2. And this precisely means that the curves L_1 and L_2 are tangent at point M_0. Note that the condition of tangency is also fulfilled in the case where point M_0 is a singularity of at least one of the curves L_1 and L_2.

Thus we have found that the condition of tangency (16.10) is fulfilled both in the case where the curves L_1 and L_2 are tangent at point M_0 and in the case where M_0 is a singularity of at least one of these curves.

16.1.2. One-parameter families of plane curves. Characteristic points of the curves of a family. Families of plane curves are often encountered in geometry and physics. In geometrical optics we have sometimes to consider reflected and refracted beams (families) of rays, in mechanics, families of possible trajectories of a particle in a given field of forces, in geometry, families of tangents to curves. One of the possible ways of defining such families of curves consists

in the following. The function $F(x, y, \alpha)$ in three variables x, y, α is considered and the curve of the family is defined for each value of

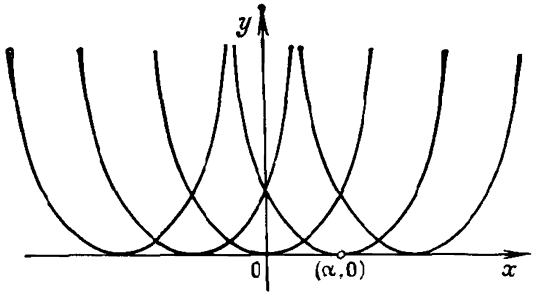


Fig. 16.1

the parameter α from the given set $\{\alpha\}$* by means of the equation

$$F(x, y, \alpha) = 0 \tag{16.12}$$

We shall say that relation (16.12) specifies a *one-parameter family of plane curves.* The parameter α will be called the parameter of the family.

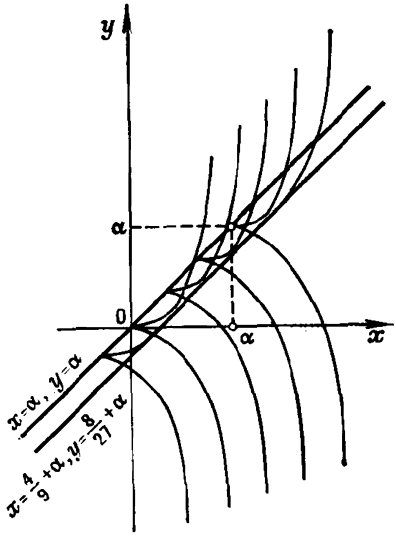


Fig. 16.2

Let us consider some examples of one-parameter families of plane curves.

1°. The equation $y - (x - \alpha)^2 = 0$ specifies the family of parabolas originating upon a displacement of the parabola $y - x^2 = 0$ along the Ox axis (Fig. 16.1).

2°. The equation $(y - \alpha)^2 - (x - \alpha)^3 = 0$ specifies the family of semi-cubic parabolas originating upon a parallel displacement of the semi-cubic parabola $y^2 = x^3$ along the bisector of the first quadrant (Fig. 16.2).

Suppose the function $F(x, y, \alpha)$ is differentiable in the domain of its definition. In this case we can introduce the notion of a *characteristic point* of the curve of the family specified by relation (16.12). Point $M(x, y)$ is called a *characteristic point* of the curve of family (16.12) corresponding to the given value α of the parameter of the family if the coordinates x and y of that point satisfy the system of equations

$$\left.\begin{aligned} F(x, y, \alpha) &= 0, \\ F'_\alpha(x, y, \alpha) &= 0. \end{aligned}\right\} \tag{16.13}$$

* The set $\{\alpha\}$ is usually some interval.

Let us discuss the geometric interpretation of the characteristic point of the curve of family (16.12). For simplicity, we shall only consider the case where any two curves of the family intersect*. Assume that L_α and $L_{\alpha+\Delta\alpha}$ are two curves of family (16.12) corresponding to the values α and $\alpha+\Delta\alpha$ of the parameter (Fig. 16.3). The coordinates of point N of their intersection satisfy the following system of equations:

$$\left.\begin{aligned} F(x, y, \alpha) &= 0, \\ F(x, y, \alpha+\Delta\alpha) &= 0 \end{aligned}\right\}$$

or the equivalent system

$$\left.\begin{aligned} F(x, y, \alpha) &= 0, \\ \frac{F(x, y, \alpha+\Delta\alpha)-F(x, y, \alpha)}{\Delta\alpha} &= 0. \end{aligned}\right\}$$

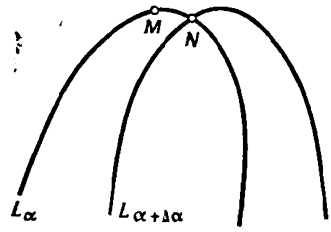


Fig. 16.3

If $\Delta\alpha \to 0$, then point N located on the curve $F(x, y, \alpha)=0$, tends, in the general case, to some point M on that curve. Since $\lim\limits_{\Delta\alpha\to 0} \frac{F(x, y, \alpha+\Delta\alpha)-F(x, y, \alpha)}{\Delta\alpha} = F'_\alpha(x, y, \alpha)$, the coordinates of point M comply with equations (16.13) and, therefore, point M is the characteristic point of the curve of the family. Thus, a characteristic point of a given curve is the point which, for $\Delta\alpha \to 0$, serves as the limit for the points of intersection of the given curve L_α and the curve $L_{\alpha+\Delta\alpha}$ lying close to it.

16.1.3. The envelope and the discriminant curve of a one-parameter family of plane curves. Suppose a one-parameter family of plane curves is specified by relation (16.12). We shall also assume that the function $F(x, y, \alpha)$ is differentiable in the domain of its definition. We shall introduce the notion of the *envelope* of the family of curves (16.12).

The envelope of a one-parameter family of curves (16.12) *is the curve O which* (1) *at each its point is tangent to only one curve of family* (16.12), (2) *at different points is tangent to different curves of that family.* Visual geometric considerations imply that the envelope is tangent to the curves of the family at the characteristic points of those curves and, therefore, under certain conditions, can be regarded as the locus of the characteristic points of the curves of the family. In fact let M be the point of tangency of the envelope O and the curve L_α of the family corresponding to the value α of the parameter of the family (Fig. 16.4), P be the point of tangency of the envelope O and the curve $L_{\alpha+\Delta\alpha}$ of the family corresponding to the value $\alpha+\Delta\alpha$ of the parameter, and N be the point of intersection of the curves L_α

* There exist families of curves whose any two curves do not intersect. Cited as an example may be a family of the parabolas specified by the equation $y - (x-\alpha)^3 = 0$.

and $L_{\alpha+\Delta\alpha}$. It can be seen from the figure that as $\Delta\alpha \to 0$, points P and N tend to point M, that is, the envelope O is tangent to the curve L_α precisely at the characteristic point M of that curve.

We shall use the term the *discriminant curve* of the family of curves (16.12) for the locus of the characteristic points of the curves of that family. Let us find out under what conditions the discriminant curve is the envelope. Let us first prove the following lemma.

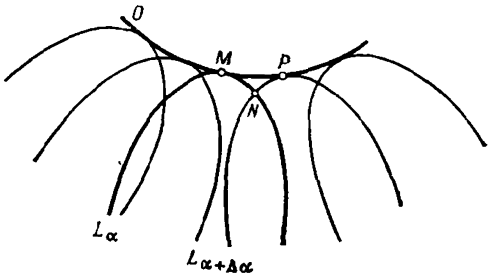


Fig. 16.4

Lemma. *Assume that $M_0(x_0, y_0)$ is a characteristic point of the family $F(x, y, \alpha) = 0$ corresponding to the value α_0 of the parameter of the family. Next assume that the functions $F(x, y, \alpha)$ and $F'_\alpha(x, y, \alpha)$ are differentiable in some neighbourhood of point (x_0, y_0, α_0) and that the partial derivatives of these functions with respect to x and y are continuous at point (x_0, y_0, α_0) itself. Then, if the Jacobian $\frac{D(F, F'_\alpha)}{D(x, y)}$ is nonzero at point (x_0, y_0, α_0), the discriminant curve passing through M_0 can be specified, in some neighbourhood of that point, by the parametric equations $x = \varphi(\alpha)$ and $y = \psi(\alpha)$, where $\varphi(\alpha)$ and $\psi(\alpha)$ are functions differentiable in some neighbourhood α_0.*

Proof. Since point $M_0(x_0, y_0)$ is a characteristic point of the curve of the family corresponding to the value α_0 of the parameter of the family, $F(x_0, y_0, \alpha_0) = 0$ and $F'_\alpha(x_0, y_0, \alpha_0) = 0$ and, therefore, by the conditions of the lemma, the system of equations (16.13) specifying the characteristic points of the curves of the family complies with all the requirements of Theorem 15.2 on the solvability of the system of equations with respect to x and y. Consequently, two functions

$$x = \varphi(\alpha), \quad y = \psi(\alpha) \tag{16.14}$$

are defined in some neighbourhood of point α_0 which are the unique continuous and differentiable solution of system (16.13). The curve specified by parametric equations (16.14) consists of characteristic points and is, therefore, the discriminant curve of the family passing

through point M_0. Note that since the solution of system (16.13) is unique, various points of the discriminant curve specified by the parametric equations (16.14) are the characteristic points of various curves of family (16.12). We shall now indicate the addition conditions under which the discriminant curve passing through point M_0 is an *envelope* in a certain neighbourhood of that point.

Theorem 16.1. *Suppose that besides the conditions formulated in the lemma, the following conditions are also satisfied:* (1) *the derivatives* F'_x, F'_y, $F''_{\alpha x}$, $F''_{\alpha y}$, $F''_{\alpha\alpha}$ *are continuous in some neighbourhood of point* (x_0, y_0, α_0); (2) *the relations* $F'^2_x + F'^2_y \neq 0$, $F''_{\alpha\alpha} \neq 0$ *are complied with at point* (x_0, y_0, α_0). *Then the discriminant curve passing through the characteristic point* $M_0(x_0, y_0)$ *is the envelope of the family of curves in question in some neighbourhood of that point.*

Proof. We have verified that under the conditions we have formulated, a certain portion of the discriminant curve passing through point M_0 can be specified by parametric equations (16.14) which are a solution of system (16.13). Let us substitute this solution into equations (16.13) and differentiate the resulting identities $F(x, y, \alpha) \equiv 0$ and $F'_\alpha(x, y, \alpha) \equiv 0$ with respect to α. We obtain

$$\left.\begin{aligned} F'_x \frac{dx}{d\alpha} + F'_y \frac{dy}{d\alpha} + F'_\alpha \equiv 0, \\ F''_{\alpha x} \frac{dx}{d\alpha} + F''_{\alpha y} \frac{dy}{d\alpha} + F''_{\alpha\alpha} \equiv 0. \end{aligned}\right\} \tag{16.15}$$

Since $F'_\alpha \equiv 0$, the first relation (16.15) assumes the form

$$F'_x \frac{dx}{d\alpha} + F'_y \frac{dy}{d\alpha} \equiv 0.$$

We see that at each point of the discriminant curve in question the condition of tangency (16.11) of that curve and the corresponding curve of the family is fulfilled. Consequently, by virtue of Remark 2 in 16.1.1, the proof of the theorem will be complete if we establish that every characteristic point of the curve of the family situated in some neighbourhood of point M_0 and every point of the discriminant curve in that neighbourhood are regular. By the hypothesis, at point (x_0, y_0, α_0) there holds a relation $F'^2_x + F'^2_y \neq 0$ which by virtue of continuity of the partial derivatives F'_x and F'_y will also be satisfied in some neighbourhood of the indicated point. Consequently, in this neighbourhood, all the characteristic points of the curves of the family are regular. It follows from the relation $F''_{\alpha\alpha} \neq 0$ (valid, because of the continuity of this derivative, in some neighbourhood of point (x_0, y_0, α_0)) and from the second identity (16.15) that the derivatives $\frac{dx}{d\alpha}$ and $\frac{dy}{d\alpha}$ do not simultaneously vanish in the indicated

neighbourhood*. Thus, all the points of the discriminant curve are regular in some neighbourhood of M_0. It also follows from the lemma we have just proved that various points of a discriminant curve are characteristic points of various curves of the family. We have proved the theorem.

Remark 1. The argument used in the proof of the theorem shows that in the case where only the conditions of the lemma are fulfilled, the condition of tangency of that curve and a curve of the family is complied with at every point of the discriminant curve. We have pointed out (see Remark 2 in 16.1.1) that the condition of tangency is also fulfilled when the common point of two curves is a singular point of at least one of them. Hence it follows that the discriminant curve can be a locus of singular points of the curves of the family (if the condition $F'^2_x + F'^2_y = 0$ is fulfilled at each characteristic point). Note that the discriminant curve itself may have singularities (if $dx/d\alpha$ and $dy/d\alpha$ are simultaneously zero).

Remark 2. Theorem 16.1 can be given the following geometric interpretation. *If neither the curves of the family nor the discriminant curve possess singularities, then the discriminant curve in question is an envelope.*

Let us consider some examples.

1°. Find the discriminant curve of the family $y - (x - \alpha)^2 = 0$.

We have $F(x, y, \alpha) = y - (x - \alpha)^2$, $F'_\alpha(x, y, \alpha) = 2(x - \alpha)$. Thus, system (16.13) has the form

$$y - (x - \alpha)^2 = 0, \quad 2(x - \alpha) = 0.$$

It follows from this that the characteristic points have the coordinates $(\alpha, 0)$ (see Fig. 16.1), and, therefore, the discriminant curve is specified by the parametric equations

$$x = \alpha, \quad y = 0.$$

We also have $F'_x = -2(x - \alpha)$, $F'_y = 1$. At the points on the discriminant curve $F'^2_x + F'^2_y = 1$. In addition, $F''_{\alpha\alpha} = -2 \neq 0$. Thus we see that the discriminant curve, the Ox axis, is an envelope.

2°. Find the discriminant curve of the family $(y - \alpha)^2 - (x - \alpha)^3 = 0$. We have $F(x, y, \alpha) = (y - \alpha)^2 - (x - \alpha)^3$, $F'_\alpha(x, y, \alpha) = -2(y - \alpha) + 3(x - \alpha)^2$. System (16.13) is of the form

$$(y - \alpha)^2 - (x - \alpha)^3 = 0, \quad -2(y - \alpha) + 3(x - \alpha)^2 = 0.$$

* The continuity of these derivatives follows directly from the continuity of the derivatives F'_x, F'_y, $F''_{\alpha x}$, $F''_{\alpha y}$ and $F''_{\alpha\alpha}$ and from relations (16.15) from which the derivatives $\frac{dx}{d\alpha}$ and $\frac{dy}{d\alpha}$ can be found by algebraic means.

It follows that the discriminant curve constitutes two straight lines determined by the parametric equations

$$x=\alpha,\quad y=\alpha \text{ and } x=\frac{4}{9}+\alpha,\quad y=\frac{8}{27}+\alpha$$

(see Fig. 16.2). It is easy to verify that the condition $F'^2_\alpha + F'^2_y = 0$ is fulfilled at the points of the straight line $x = \alpha$ and $y = \alpha$, i.e. this portion of the discriminant curve is the locus of singular points of the curves of the family. The reader can easily ascertain that the line $x = 4/9 + \alpha$, $y = 8/27 + \alpha$ is the envelope of the family of curves being considered.

16.1.4. The envelope and the discriminant surface of a one-parameter family of surfaces. Let us consider a one-parameter family of surfaces specified by the equation

$$F(x, y, z, \alpha) = 0. \tag{16.16}$$

We assume that the function $F(x, y, z, \alpha)$ is differentiable in the domain of its definition. The line L on the surface of the family (16.16) corresponding to the value α of the parameter of the family is called a *characteristic* if the coordinates of the points of that line satisfy the system of equations

$$\left.\begin{aligned} F(x, y, z, \alpha)&=0,\\ F'_\alpha(x, y, z, \alpha)&=0. \end{aligned}\right\} \tag{16.17}$$

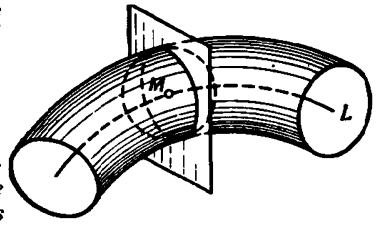


Fig. 16.5

The locus of characteristics is called a *discriminant surface* of family (16.17). The *envelope of a one-parameter family of surfaces* (16.16) *is the surface O which is in contact with all the surfaces of the family.*

The following assertion can be proved.

If neither the surfaces of the family nor the discriminant surface possess singularities, the indicated discriminant surface is the envelope.

Note that a discriminant surface may be a locus of singular points of the surfaces of the family and may possess singularities itself.

Let us discuss the following example. Find the envelope of a family of spheres of constant radius R whose centres are at the points of a given curve L* (shown in Fig. 16.5 is one such sphere with centre at point M of the curve L), specified by the equations

$$x = \varphi(\alpha),\quad y = \psi(\alpha),\quad z = \chi(\alpha).$$

The one-parameter family of spheres in question is specified by the following equation:

$$[x - \varphi(\alpha)]^2 + [y - \psi(\alpha)]^2 + [z - \chi(\alpha)]^2 - R^2 = 0. \tag{16.18}$$

The characteristics of the indicated family of spheres are determined from equation (16.18) and from the equation

$$[x - \varphi(\alpha)]\,\varphi'(\alpha) + [y - \psi(\alpha)]\,\psi'(\alpha) + [z - \chi(\alpha)]\,\chi'(\alpha) = 0. \tag{16.19}$$

Equation (16.19) is a plane passing through the centre M of the sphere at right angles to the tangent to the curve L. Therefore, the characteristics are the circles constituting the lines of intersection of the spheres being considered and planes (16.19) (see Fig. 16.5). Note that if L is a circle then the envelope is a torus.

* The envelopes of such families of spheres are called *canal surfaces.*

16.2. CONTACT OF PLANE CURVES

16.2.1. The notion of the order of contact of plane curves. Suppose two curves L_1 and L_2 are in contact at some point M_0^* (Fig. 16.6). Also suppose that M is an arbitrary point on a common tangent to the curves L_1 and L_2, and M_1 and M_2 are the points of intersection of a perpendicular erected at point M^{**} and the curves L_1 and L_2 respectively (see Fig. 16.6). *We shall say that at point M_0 two curves L_1 and L_2 have contact of order n, if there exists a nonzero limit*

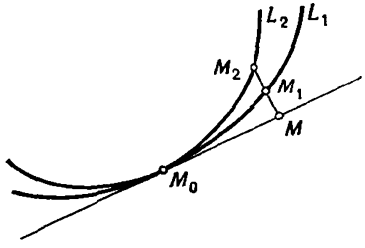


Fig. 16.6

$$\lim_{M \to M_0} \frac{|M_1M_2|}{|MM_0|^{n+1}}. \tag{16.20}$$

(It this case $|M_1M_2|$ denotes the length of the line segment M_1M_2 and $|MM_0|$ denotes the length of the line segment MM_0.)

Remark 1. If limit (16.20) is equal to zero, then we say that the curves L_1 and L_2 touch each other or have contact of the order *higher than* n.

Remark 2. If at point M_0 two curves L_1 and L_2 have contact of the order higher than *any* n, then we say that at this point they have *contact of an infinite order.*

Examples. 1°. The curves which are graphs of the functions $y = x^2$ and $y = 3x^2$ touch each other at the origin O, the Ox axis being their common tangent. Taking point M with abscissa x on the x-axis, we obtain $|OM| = |x|$, and $|M_1M_2| = |3x^2 - x^2| = 2x^2$.

Since

$$\lim_{M \to 0} \frac{|M_1M_2|}{|OM|^2} = \lim_{x \to 0} \frac{2x^2}{|x|^2} = 2 \neq 0,$$

the curves under consideration have contact of order unity at point O.

2°. Let us now consider two curves, L_1 and L_2, the first of which coincides with the x-axis and the other is a graph of the function

$$y = \begin{cases} e^{-\frac{1}{|x|}} & \text{at } x \neq 0, \\ 0 & \text{at } x = 0. \end{cases}$$

Let us verify that curves have contact of an infinite order at the origin O. Since the common tangent at point O

* That is, pass through M_0 and have coinciding tangents at that point.

** In this case it is assumed that if point M is sufficiently close to M_0, then the perpendicular erected at point M to the tangent intersects each curve L_1 and L_2 at only one point.

is the Ox axis, we can take point M with abscissa x on that axis and obtain $|OM| = |x|$ and $|M_1M_2| = e^{-\frac{1}{|x|}}$. It is sufficient to prove that for any n the limit $\lim\limits_{M\to 0} \frac{|M_1M_2|}{|OM|^{n+1}} = \lim\limits_{x\to 0} \frac{e^{-\frac{1}{|x|}}}{|x|^{n+1}}$ is equal to zero. Assuming $t = \frac{1}{|x|}$, we reduce this limit to the limit $\lim\limits_{t\to+\infty} \frac{t^{n+1}}{e^t}$. We have proved at the end of 8.12.2 that the last limit is equal to zero.

16.2.2. The order of contact of curves which are graphs of functions. Let us assume that two curves, L_1 and L_2, are the graphs of the functions $y = f_1(x)$ and $y = f_2(x)$ respectively. Let us assume, further, that the curves are in contact at some point $M_0(x_0, f_1(x_0))$, point M_0 being regular for each of these curves*. We shall prove that under these assumptions, the definition of the order of contact of the curves L_1 and L_2 given in 16.2.1 can be replaced by an equivalent definition, more convenient for applications.

Let Δx be an arbitrary increment of the argument at point x_0 and $x = x_0 + \Delta x$. *We shall say that at point* $M_0(x_0, f_1(x_0))$ *the curves* L_1 *and* L_2 *have contact of order n if there is a nonzero limit*

$$\lim_{\Delta x\to 0} \frac{|f_2(x_0+\Delta x) - f_1(x_0+\Delta x)|}{|\Delta x|^{n+1}} = \lim_{x\to x_0} \frac{|f_2(x) - f_1(x)|}{|x-x_0|^{n+1}}. \qquad (16.21)$$

To speak of the order of contact of the curves L_1 and L_2 in the sense of the definition given in 16.2.1, we have first to prove that in some neighbourhood of M_0 these curves are projected in a *one-to-one* manner on their common tangent. Lemma 1 given below proves exactly this fact. The remaining part of 16.2.2 is devoted to the proof of another two lemmas which directly yield the equivalence of the two definitions of the order of contact of the curves L_1 and L_2.

Lemma 1. *If point* $M_0(x_0, f(x_0))$ *is a regular point of the curve L which is a graph of the function* $y = f(x)$, *then in some neighbourhood of* M_0 *the curve L is projected in a one-to-one manner on its tangent.*

Proof. Let us note, first of all, that the existence of a tangent to the curve L at point M_0 follows from the existence of a derivative $f'(x_0)$. We shall now pass to a new system of rectangular Cartesian coordinates XY, positioning the origin at the point of tangency M_0 and directing the X-axis along the tangent at M_0 (Fig. 16.7). If we denote by α the angle between the x- and X-axes, it becomes evident that the new system of coordinates is obtained from the old one by shifting the origin to point $M_0(x_0, f(x_0))$ and turning the axes through the angle α. Using the known formulas for the transformation of coordinates upon the displacement and rotation of the axes, we shall obtain the following expres-

* The definition of a regular point of a curve specified by the equation $F(x, y) = 0$ was given in 16.1.1. In a particular case, for the curve specified by the equation $y = f(x)$ point $M_0(x_0, f(x_0))$ will be regular if the derivative $f'(x)$ is continuous only at point x_0.

sion for the new coordinates X, Y of the points of the curve L in terms of the old coordinates x and $y = f(x)$:

$$\left.\begin{aligned} X &= (x - x_0)\cos\alpha + [f(x) - f(x_0)]\sin\alpha, \\ Y &= -(x - x_0)\sin\alpha + [f(x) - f(x_0)]\cos\alpha. \end{aligned}\right\} \tag{16.22}$$

We shall attain our aim if we show that using equations (16.22) we can express Y as a function of X (this will precisely mean that the portion of the curve L adjacent to M_0 can be projected in a one-to-one manner on the X-axis, i.e. on the tangent). To do that, it is sufficient to prove that in the neighbourhood of

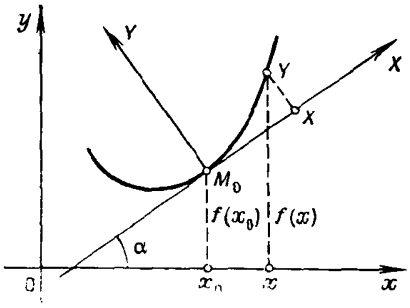


Fig. 16.7

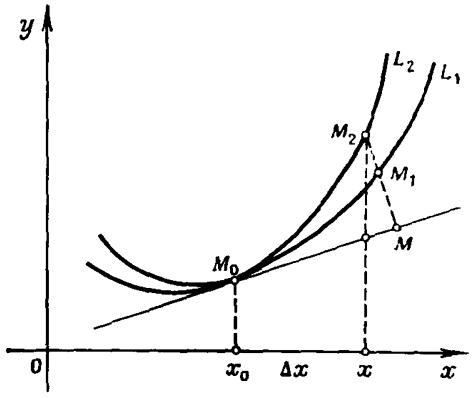


Fig. 16.8

M_0 the first equation (16.22) has a unique solution with respect to x. (Indeed, expressing x in terms of X from the first equation (16.22) and substituting the resulting expression into the second equation (16.22), we shall determine Y as a function of X.) In accordance with Theorem 15.1, it is sufficient to prove that the derivative with respect to x of the right-hand side of the first equation (16.22) is nonzero at point x_0 and is continuous at that point. But this is obvious since the indicated derivative has the form $\cos\alpha + f'(x)\sin\alpha$ and is equal to $\dfrac{1}{\cos\alpha}$ at point x_0 at which $f'(x_0) = \tan\alpha$. Since $\alpha \neq \dfrac{\pi}{2}$, it follows that $\cos\alpha \neq 0$. We have proved the lemma.

Assume that the curves L_1 and L_2, which are graphs of the functions $y = f_1(x)$ and $y = f_2(x)$, touch each other at point $M_0(x_0, f_1(x_0))$, which is regular for each of these curves, and $x = x_0 + \Delta x$. Further assume that a perpendicular is dropped from point $M_2(x, f_2(x))$ onto the tangent line at M_0 (Fig. 16.8), point M_1 is the point of intersection of that perpendicular with the curve L_1*, and M is the point of intersection of the perpendicular with the tangent. Then the following assertions are true.

Lemma 2. *There exists a nonzero limit*

$$\lim_{\Delta x \to 0} \frac{|f_2(x) - f_1(x)|}{|M_1 M_2|}. \tag{16.23}$$

Lemma 3. *There exists a nonzero limit*

$$\lim_{\Delta x \to 0} \frac{|M_0 M|}{|\Delta x|}. \tag{16.24}$$

* By virtue of Lemma 1, this point will be unique for a sufficiently small Δx.

Note that it follows from Lemmas 2 and 3 that the nonzero limit (16.21) exists if and only if there exists a nonzero limit (16.20), and this precisely signifies the equivalence of the two definitions of the order of contact. Let us now prove Lemmas 2 and 3.

Proof of Lemma 2. For the sake of simplicity, we shall designate the distance M_1M_2 by ρ, and shall assume for definiteness that the curves L_1 and L_2 are

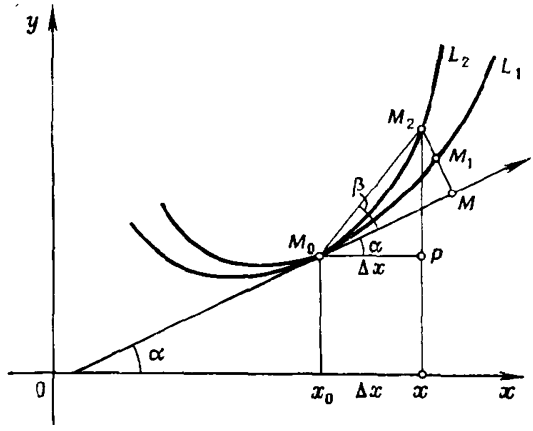


Fig. 16.9

positioned as shown in Fig. 16.8. If X and Y are the coordinates of point M_2 and α is the slope of the tangent to the x-axis at M_0, then it is evident that

$$X = x + \rho \sin \alpha, \quad Y = f_2(x) - \rho \cos \alpha.$$

Since point M_1 lies on the curve L_1, it follows that $Y = f_1(x)$, or

$$f_2(x) - \rho \cos \alpha = f_1(x + \rho \sin \alpha).$$

Applying Lagrange's formula for the closed interval $[x, x + \rho \sin \alpha]$ to the right-hand side of the last equality, we shall have

$$f_2(x) - \rho \cos \alpha = f_1(x) + \rho \sin \alpha f'(\xi),$$

where ξ is some point lying within the indicated closed interval. Since $\rho \to 0$ for $\Delta x \to 0$ and the derivative $f'(x)$ is continuous at point x_0, the last equality can be rewritten in the form

$$f_2(x) - \rho \cos \alpha = f_1(x) + \rho \sin \alpha [f'(x_0) + \varepsilon], \tag{16.25}$$

where $\varepsilon \to 0$ for $\Delta x \to 0$. Taking into account that $f'(x_0) = \tan \alpha$, we shall obtain from (16.25)

$$\frac{f_2(x) - f_1(x)}{\rho} = \frac{1}{\cos \alpha} + \varepsilon \sin \alpha.$$

Lemma 2 is proved.

Proof of Lemma 3. Let us designate by β the angle between the chord M_0M_2 and the tangent M_0M (Fig. 16.9). Evidently, $\beta \to 0$ for $\Delta x \to 0$. We obtain from the right-angled triangles M_0M_2P and M_0M_2M (here M_0P is parallel to the x-axis) that

$$|M_0M_2| = \frac{|\Delta x|}{\cos(\alpha + \beta)} = \frac{|M_0M|}{\cos \beta}.$$

It is evident from the last formula that

$$\lim_{\Delta x \to 0} \frac{|M_0 M|}{|\Delta x|} = \lim_{\Delta x \to 0} \frac{\cos \beta}{\cos (\alpha + \beta)} = \frac{1}{\cos \alpha}.$$

We have proved Lemma 3.

16.2.3. Sufficient conditions of contact of order n. Assume, as before, that two curves L_1 and L_2, which are graphs of the functions $f_1(x)$ and $f_2(x)$ touch each other at $M_0(x_0, f_1(x_0))$. The following statement is true.

Theorem 16.2. *Suppose that the functions $y = f_1(x)$ and $y = f_2(x)$ are $n+1$ times differentiable in some neighbourhood of point x_0, the derivatives of order $n+1$ being continuous at point x_0 itself. Then, if the relations*

$$f_1(x_0) = f_2(x_0),\ f_1'(x_0) = f_2'(x_0),\ \ldots,\ f_1^{(n)}(x_0) = f_2^{(n)}(x_0),$$
$$f_1^{(n+1)}(x_0) \neq f_2^{(n+1)}(x_0), \tag{16.26}$$

are fulfilled at point x_0, then at point $M_0(x_0, f_1(x_0))$ the curves L_1 and L_2 have contact of order n.

Proof. Assume $F(x) = f_2(x) - f_1(x)$. It is sufficient to prove that there exists a nonzero limit

$$\lim_{x \to x_0} \frac{|F(x)|}{|x - x_0|^{n+1}}.$$

Since $F(x_0) = F'(x_0) = \ldots = F^{(n)}(x_0) = 0$, $F^{(n+1)}(x_0) \neq 0$ by virtue of (16.26) we can write Taylor's formula with the remainder in the form of Lagrange for the function $F(x)$ and obtain

$$F(x) = \frac{1}{(n+1)!} F^{(n+1)}(x_0 + \theta(x - x_0))(x - x_0)^{n+1}, \quad 0 < \theta < 1. \tag{16.27}$$

The derivative of the order $n+1$ being continuous at point x_0, we have

$$F^{(n+1)}(x_0 + \theta(x - x_0)) = F^{(n+1)}(x_0) + \varepsilon, \tag{16.28}$$

where $\varepsilon \to 0$ as $x \to x_0$. We obtain from relations (16.27) and (16.28)

$$\lim_{x \to x_0} \frac{|F(x)|}{|x - x_0|^{n+1}} = \frac{1}{(n+1)!} |F^{n+1}(x_0)| \neq 0.$$

The proof of the theorem is complete.

Remark 1. If all the conditions of Theorem 16.2 are fulfilled, with the exception, may be, of the condition $f_1^{(n+1)}(x_0) \neq f_2^{(n+1)}(x_0)$, we can state that at point M_0 the curves L_1 and L_2 have contact of the order *not lower* than n.

Remark 2. Let us find the order of contact of the curve L, which is a graph of the function $y = f(x)$, with its tangent T at point

$M_0(x_0, f(x_0))$. We shall assume that the function $f(x)$ is thrice differentiable in the neighbourhood of point x_0 and that its third derivative is continuous at point x_0. Recall that the tangent T is a graph of the linear function $y = \varphi(x) = f'(x_0)(x - x_0) + f(x_0)$. Since $\varphi(x_0) = f(x_0)$, $\varphi'(x_0) = f'(x_0)$ and $\varphi''(x_0) = 0$, it follows that in the case $f''(x_0) \neq 0$ the curve L has contact of order $n = 1$ with its tangent T, and in the case $f''(x_0) = 0$ the order of contact is not lower than two.

16.2.4. An osculating circle. Assume that $M_0(x_0, y_0)$ is a point of the curve L which is a graph of the function $y = f(x)$ possessing a continuous third derivative at point x_0. We can draw through point M_0 infinitely many circles touching the curve L at that point. It is easy to verify that a portion of every one of these circles lying in some neighbourhood of M_0 is a graph of the function of the form $y = y(x)$. We can, therefore, speak of the order of contact of the curve L with any of these circles at their common point M_0. That of the circles which has *contact* with the curve L *of the order not lower than two* is known as an *osculating circle for the curve* L *at point* M_0. The following assertion states the sufficient conditions of the existence of an osculating circle for the curve L at point M_0.

Theorem 16.3. *Let the curve L be a graph of the function $y = f(x)$, the function $f(x)$ having at point x_0 a nonzero second derivative and a continuous third derivative. Then, there exists an osculating circle for the curve L at point $M_0(x_0, f(x_0))$.*

Proof. We shall search for an equation of an osculating circle in the form

$$(x - a)^2 + (y - b)^2 = \rho^2, \tag{16.29}$$

where a, b and ρ are constants which we have to determine. Let us solve equation (16.29) with respect to y and substitute the resulting solution $y = y(x)$ into the left-hand side of (16.29). Then we can regard (16.29) as an identity with respect to x (assuming, of course, that $y = y(x)$). We shall differentiate identity (16.29) twice with respect to x and put a requirement that in the relations obtained and in relation (16.29) itself the values y_0, y_0', y_0'' of the function $y = y(x)$ and of its first two derivatives at point x_0 be equal to $f(x_0)$, $f'(x_0)$, $f''(x_0)$ respectively:

$$y_0 = f(x_0),\ y_0' = f'(x_0),\ y_0'' = f''(x_0).$$

We shall thus obtain the following system of equations with respect to a, b and ρ:

$$(x_0 - a)^2 + (y_0 - b)^2 = \rho^2,$$

$$(x_0 - a) + (y_0 - b)\, y_0' = 0,$$

$$1 + (y_0')^2 + (y_0 - b)\, y_0'' = 0.$$

Under the condition $y_0'' = f''(x_0) \neq 0$, $\rho > 0$, this system has the unique solution

$$a = x_0 - \frac{1+y_0'^2}{y_0''} y_0', \quad b = y_0 + \frac{1+y_0'^2}{y_0''}, \quad \rho = \frac{(1+y_0'^2)^{3/2}}{|y_0''|}. \tag{16.30}$$

It follows from the reasoning we have presented that for the circle the coordinates (a, b) of whose centre and the radius ρ are determined from formulas (16.30) and for the curve L all the conditions of Remark 1 to Theorem 16.2 are fulfilled.

Remark. If in the hypothesis of Theorem 16.3, the requirement $f''(x_0) \neq 0$ is replaced by the contrary requirement $f''(x_0) = 0$, there will be no osculating circle for the curve L at point M_0.

However, by virtue of Remark 2 to Theorem 16.2, in that case the curve L and the tangent to it at point M_0 have contact of the order not lower than two at that point. Thus it can be taken as a fact that in the indicated case the osculating circle at point M_0 degenerates into a straight line.

16.3. THE CURVATURE OF A PLANE CURVE

16.3.1. The notion of the curvature of a plane curve. Suppose the curve L is specified by the parametric equations $x = \varphi(t), y = \psi(t)$. We shall consider the conditions stated in 11.1 to be fulfilled for the curve L*. Then we can choose a positive direction at any point M of the curve L. Let us agree to take for the positive direction at a given point M of the curve L the direction in which point M will move upon an increase in the parameter t.

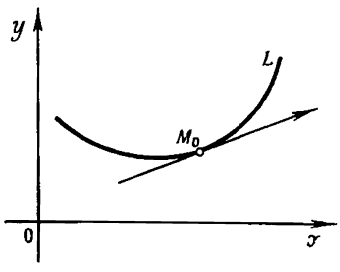


Fig. 16.10

The next step is to fix, on the curve L, a certain point M_0 corresponding to the value of the parameter t_0 and assume that this point is a *regular point* of the curve L.

Then the portion of the curve L adjacent to point M_0 is a graph of the function of the form $y = f(x)$ or of the form $x = f_1(y)$ and, besides, there is a tangent to the curve L at point M_0. We shall introduce a positive direction along this tangent corresponding to the positive direction along the curve L at point M_0**.

Everywhere in what follows we shall only consider a directed tangent line. In Fig. 16.10 the direction of the tangent line is shown by an arrow.

* That is, we shall consider the curve L to be free from points of self-intersection and of parts of self-superimposition.

** We shall agree to consider as a positive direction along the tangent the direction in which a point, which is the projection of a point of the curve L on the tangent, will move upon an increase in the parameter.

Let us assume now that *all the points of the curve L* lying in some neighbourhood of point M_0 we have fixed *are regular*. Suppose M is one of these points. Let us now introduce the notion of an *angle of contingence* of a portion M_0M of the curve. We shall assume for definiteness that point M corresponds to the greater value of the parameter than point M_0.

The *angle of contingence of a portion* M_0M *of the curve* is the angle between the directed tangents to the curve L at points M and M_0 taken with the plus sign if the shortest way to bring the tangent

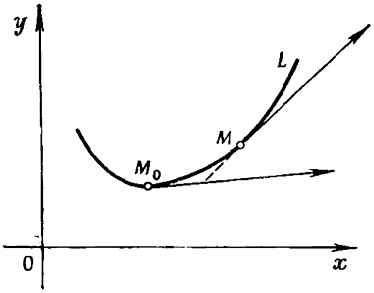


Fig. 16.11

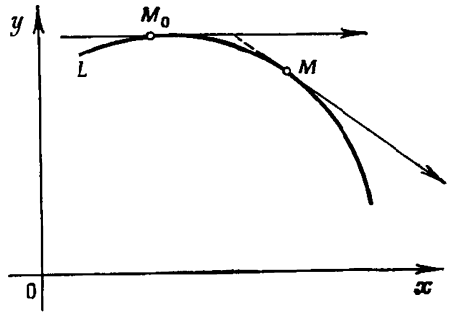


Fig. 16.12

at point M_0 into coincidence with the tangent at point M is to rotate it counterclockwise and with the minus sign otherwise.

Fig. 16.11 shows the portion M_0M having a positive angle of contingence and Fig. 16.12 shows the portion M_0M having a negative angle of contingence.

Next we shall introduce the notion of the *average curvature of the portion* M_0M *of the curve. The average curvature of the portion* M_0M *of the curve is the ratio of the angle of contingence of that portion to its length.*

Since point M_0 of the curve L is assumed to be fixed, the average curvature of the portion M_0M is a function of point M or a function of the parameter t. We shall designate this function by $\chi_{M_0}(M)$ or $\chi_{M_0}(t)$.

It is natural to consider here the problem concerning the limit of this function in the case where point M tends to M_0 along the curve L (or, which is the same, where the parameter t tends to t_0).

Definition. *The limiting value of the average curvature of the portion* M_0M *of the curve in the case where point* M *tends to* M_0 *along the curve is called the* **curvature** *of the curve* L *at the given point* M_0 *and is designated as* $k(M_0)$.

Thus, according to the definition,

$$k(M_0) = \lim_{M \to M_0} \chi_{M_0}(M) = \lim_{t \to t_0} \chi_{M_0}(t).$$

The reader can make sure that (1) both the average curvature of any portion of a straight line and the curvature at any point of that line are equal to zero, (2) both the average curvature of any portion of a circle of radius R and the curvature at any point of that circle are equal to $\frac{1}{R}$.

Let us establish the formula to calculate the curvature at any point of an arbitrary curve L.

16.3.2. Formula for calculating curvature. Suppose that the curve L is specified by the parametric equations $x = \varphi(t)$, $y = \psi(t)$, and M_0 is some fixed point of this curve corresponding to the value t_0 of the parameter. Assume that all the points of the curve L from some neighbourhood of M_0 are regular and that the functions $\varphi(t)$ and $\psi(t)$ possess the second derivatives at point t_0. Under these assumptions we shall establish a general formula for calculating the curvature of the curve L at point M_0.

Assume that $\dot{x}$ and $\ddot{x}$ are the values of the first and the second derivative of the function $x = \varphi(t)$ at point t_0, and $\dot{x} + \dot{\Delta x}$ is the value of the first derivative of that function at point $t_0 + \Delta t$ (Δt being an arbitrary increment of the parameter t). Thus, $\dot{\Delta x}$ is the increment of the first derivative of the function $x = \varphi(t)$. Let $\dot{y}$, $\ddot{y}$ and $\dot{y} + \dot{\Delta y}$ be the corresponding values of the derivatives of the function $y = \psi(t)$.

If we assume point M_0, corresponding to the value t_0 of the parameter, to be fixed, and point M to correspond to the value $t = t_0 + \Delta t$ of the parameter then the angle of contingence of the portion M_0M and the length of this portion can be regarded as *functions of the argument* Δt. The functions will be denoted by $\alpha(\Delta t)$ and $l(\Delta t)$ respectively.

In accordance with the definition, the curvature $k(M_0)$ at point M_0 of the curve is equal to the limiting value

$$k(M_0) = \lim_{\Delta t \to 0} \frac{\alpha(\Delta t)}{l(\Delta t)} \tag{16.31}$$

Let us prove that the limiting value (16.31) exists and calculate it.

Let us denote by φ_0 and φ the slopes to the x-axis of the tangents to the curve L drawn through points M_0 and M respectively (Fig. 16.13). Then, evidently, for any position of points M_0 and M, the following relation will be valid for the angle of contingence $\alpha(\Delta t)$:

$$\alpha(\Delta t) = \varphi - \varphi_0.$$

This relation yields

$$\tan \alpha(\Delta t) = \frac{\tan \varphi - \tan \varphi_0}{1 + \tan \varphi \tan \varphi_0}. \tag{16.32}$$

Proceeding from the geometrical meaning of a derivative (see 5.1.4) and from the expression for the derivative of a function specified parametrically (see 5.11), we can write

$$\tan \varphi_0 = \frac{\dot{y}}{\dot{x}}, \quad \tan \varphi = \frac{\dot{y}+\Delta\dot{y}}{\dot{x}+\Delta\dot{x}}.$$

Thus, formula (16.32) can be rewritten in the form

$$\tan \alpha(\Delta t) = \frac{\dfrac{\dot{y}+\Delta\dot{y}}{\dot{x}+\Delta\dot{x}} - \dfrac{\dot{y}}{\dot{x}}}{1+\dfrac{\dot{y}(\dot{y}+\Delta\dot{y})}{\dot{x}(\dot{x}+\Delta\dot{x})}} = \frac{\dot{x}\cdot\Delta\dot{y}-\dot{y}\cdot\Delta\dot{x}}{\dot{x}(\dot{x}+\Delta\dot{x})+\dot{y}(\dot{y}+\Delta\dot{y})}. \tag{16.33}$$

Since the functions $x = \varphi(t)$ and $y = \psi(t)$ possess the second deriv-

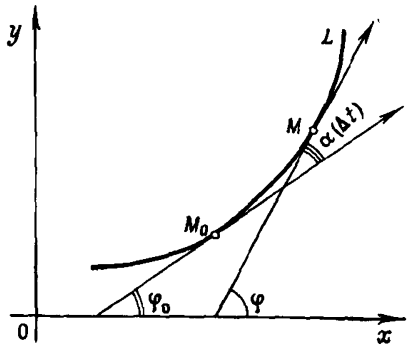


Fig. 16.13

atives at point t_0, the first derivatives of these functions are continuous at point t_0 and, hence,

$$\lim_{\Delta t \to 0} \Delta\dot{y} = 0, \quad \lim_{\Delta t \to 0} \Delta\dot{x} = 0.$$

But then, by virtue of equality (16.33),

$$\lim_{\Delta t \to 0} \tan \alpha(\Delta t) = 0.$$

It follows from the last equality and from the continuity of the arctangent that

$$\lim_{\Delta t \to 0} \alpha(\Delta t) = 0.$$

Therefore, the following equality holds true:

$$\lim_{\Delta t \to 0} \frac{\alpha(\Delta t)}{\tan \alpha(\Delta t)} = \lim_{\Delta t \to 0} \left[\cos \alpha(\Delta t) \frac{\alpha(\Delta t)}{\sin \alpha(\Delta t)}\right] = 1. \tag{16.34}$$

Equality (16.34) and the theorem on the limiting value of a product reduce the problem of computing the limiting value (16.31) to that of computing the following limiting value:

$$k(M_0) = \lim_{\Delta t \to 0} \frac{\tan \alpha (\Delta t)}{l(\Delta t)}. \tag{16.35}$$

To calculate this limit, pay attention to the fact that the length $l(\Delta t)$ of the portion M_0M of the curve is determined by the formula

$$l(\Delta t) = \int_{t_0}^{t_0+\Delta t} \sqrt{\dot{x}^2(\tau) + \dot{y}^2(\tau)}\, d\tau.$$

Applying the mean-value formula to the integral on the right-hand side of the last equality, we obtain

$$l(\Delta t) = \Delta t \sqrt{\dot{x}^2(t^*) + \dot{y}^2(t^*)}, \tag{16.36}$$

where $t_0 \leqslant t^* \leqslant t_0 + \Delta t$.

Relations (16.35), (16.33) and (16.36) enable us to infer that the calculation of the curvature reduces to the calculation of the limiting value

$$k(M_0) = \lim_{\Delta t \to 0} \frac{\dot{x}\frac{\Delta \dot{y}}{\Delta t} - \dot{y}\frac{\Delta \dot{x}}{\Delta t}}{[\dot{x}(\dot{x}+\Delta \dot{x}) + \dot{y}(\dot{y}+\Delta \dot{y})]\sqrt{\dot{x}^2(t^*)+\dot{y}^2(t^*)}}. \tag{16.37}$$

Since the functions $x = \varphi(t)$ and $y = \psi(t)$ possess the second derivatives at point t_0, there exist limiting values

$$\lim_{\Delta t \to 0} \frac{\Delta \dot{x}}{\Delta t} = \ddot{x} \quad \text{and} \quad \lim_{\Delta t \to 0} \frac{\Delta \dot{y}}{\Delta t} = \ddot{y}. \tag{16.38}$$

Further, from the continuity of the first derivatives of the functions $x = \varphi(t)$ and $y = \psi(t)$ we infer that

$$\lim_{\Delta t \to 0} \Delta \dot{x} = 0, \quad \lim_{\Delta t \to 0} \Delta \dot{y} = 0, \quad \lim_{\Delta t \to 0} \sqrt{\dot{x}^2(t^*) + \dot{y}^2(t^*)} = \sqrt{\dot{x}^2 + \dot{y}^2}. \tag{16.39}$$

The existence of limiting values (16.38) and (16.39) imply that the limit on the right-hand side of (16.37) exists and is equal to

$$\frac{\dddot{x}\,\dot{y} - \dddot{y}\,\dot{x}}{[\dot{x}^2 + \dot{y}^2]^{3/2}}.$$

Thus we have proved that the curvature of the curve L at point M_0 exists and is specified by the formula

$$k(M_0) = \frac{\ddot{x}\,\dot{y} - \ddot{y}\,\dot{x}}{[\dot{x}^2 + \dot{y}^2]^{3/2}}. \tag{16.40}$$

Remark. Suppose we have to calculate, at a given point M_0, the curvature $k(M_0)$ of the curve L *which is a graph of the twice differentiable function* $y = f(x)$.

Putting $x = t$ and $y = f(t)$ in formula (16.40), we obtain the following formula for the sought-for curvature:

$$k(M_0) = \frac{f''(x)}{[1+(f'(x))^2]^{3/2}}.$$

Finally, if the curvature is given by the polar equation $r = r(\theta)$, where $r(\theta)$ is a twice differentiable function of the polar angle θ, then, assuming the polar angle θ to be the parameter t and taking into account that $x = r(\theta)\cos\theta$, $y = r(\theta)\sin\theta$, we obtain the following expression for the curvature:

$$k = \frac{r^2(\theta) + 2[r'_\theta(\theta)]^2 - r(\theta)\cdot r''_{\theta^2}(\theta)}{\{r^2(\theta) + [r'_\theta(\theta)]^2\}^{3/2}}.$$

Let us calculate, for example, the curvature at an arbitrary point of the catenary $y = a\cosh\frac{x}{a}$.

Since

$$1 + [f'(x)]^2 = \cosh^2\frac{x}{a} = \frac{y^2}{a^2}, \quad f''(x) = \frac{1}{a}\cosh\frac{x}{a} = \frac{y}{a^2},$$

the curvature is equal to

$$k = \frac{a}{y^2} = \frac{1}{a\cosh^2\frac{x}{a}}.$$

16.4. THE EVOLUTE AND THE INVOLUTE

16.4.1. A normal to a plane curve. Suppose a plane curve L is specified by the parametric equations

$$x = \varphi(t), \quad y = \psi(t). \tag{16.41}$$

We shall assume that the curve L has no points of self-intersection or portions of self-superimposition. In addition, we shall assume that every point of the curve L is regular*.

It is time now to introduce the notion of a normal to a curve at its given point M.

A straight line lying in the plane of the curve L and passing through point M of the curve at right angles to the tangent line to L at point M is called a normal to L at point M (Fig. 16.14).

* Note that under these conditions there is a tangent at every point of the curve L.

Let us find the equation for a normal to a curve. Let x and y be the coordinates of point M of the curve L, and X and Y, the coordinates of any point N of the normal to the curve. In accordance with the definition, the normal is perpendicular to the tangent and, there-

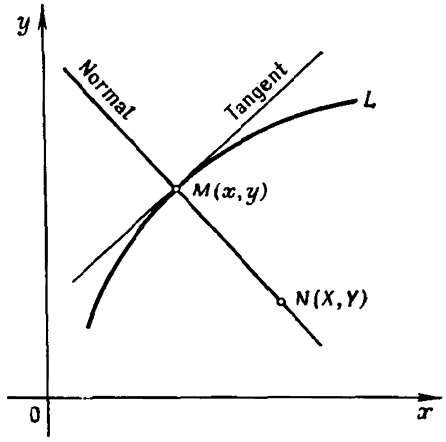


Fig. 16.14

fore, the slope k_n of the normal is connected to the slope k_t of the tangent by the relation*

$$k_n k_t = -1. \tag{16.42}$$

In the case of the parametric specification of the curve (16.41) the slope k_t of the tangent is equal to ψ'/φ', consequently we obtain from (16.42)

$$k_n = -\frac{\varphi'}{\psi'}. \tag{16.43}$$

Using the equation of a straight line with a given slope k_n, known from the course of analytic geometry, we shall obtain the following equation of a normal to the curve L:

$$Y - y = -\frac{\varphi'}{\psi'}(X - x). \tag{16.44}$$

Taking into account that $x = \varphi(t)$ and $y = \psi(t)$, we shall rewrite equation (16.44) in the form

$$\varphi'(X - \varphi) + \psi'(Y - \psi) = 0. \tag{16.45}$$

Remark. When the tangent at point M is parallel to the abscissa axis, its slope k_t is equal to zero and, therefore, relations (16.42),

* This relation can be found, for instance, in Pogorelov's *Analytic Geometry*, Mir Publishers, Moscow.

(16.43) and (16.44) are meaningless. Equation (16.45), however, is the equation of a normal in this case as well. Indeed, if $k_t = 0$ (the tangent line is parallel to the abscissa axis), we have $\psi' = 0$ while $\varphi' \neq 0$* and, consequently, relation (16.44) assumes the form $X - \varphi = 0$. But this is just the equation of a straight line perpendic-

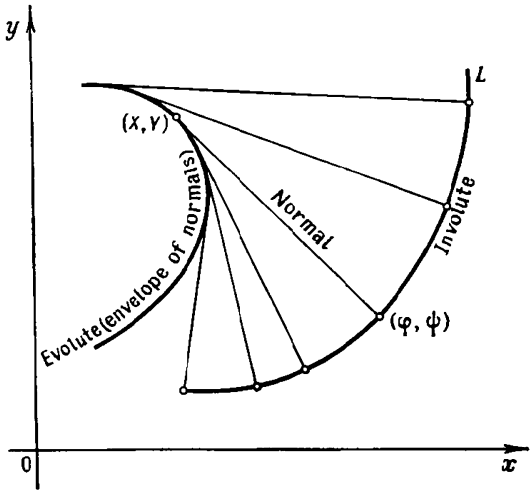


Fig. 16.15

ular to the x-axis and intercepting a segment of this straight line equal to φ. It is clear that in this case the straight line coincides with the normal at point M.

16.4.2. The evolute and the involute of a plane curve. Suppose that the curve L complies with the same conditions as in the previous section. Let us turn to equation (16.45). If t is regarded as a parameter in this equation, then it is an equation *of a one-parameter family of all the normals* of the plane curve L. The concept of a family of normals of a plane curve will be clear if you study Fig. 16.15.

Under certain conditions, a one-parameter family of normals has an envelope which is called an evolute of the curve L.

Thus, *the* **evolute** *of the plane curve L is the envelope of the one-parameter family of normals of that curve. With respect to this evolute, the curve L is called the involute.*

Let us elucidate the conditions of the existence of the evolute of the plane curve L and find its parametric equations.

Let us assume that the curve L, having no singularities, is specified by parametric equations (16.41). We shall also assume, for the sake of simplicity, that the parameter t varies on the interval (0, 1).

* Recall that $k_t = \dfrac{\psi'}{\varphi'}$ and for a regular point $\varphi'^2 + \psi'^2 \neq 0$.

Let us also suppose that the functions $\varphi(t)$ and $\psi(t)$ appearing in relations (16.41) possess continuous third derivatives on the interval (0, 1). Under these assumptions the following theorem is true.

Theorem 16.4. *Assume that at all the points of the curve L its curvature k and the derivative of the curvature* are nonzero. Then, there exists an evolute of the curve L, the parametric equations of the evolute being of the form*

$$\left.\begin{aligned} X &= \varphi - \frac{\varphi'^2+\psi'^2}{\varphi'\psi''-\varphi''\psi'}\psi', \\ Y &= \psi + \frac{\varphi'^2+\psi'^2}{\varphi'\psi''-\varphi''\psi'}\varphi'. \end{aligned}\right\} \tag{16.46}$$

Proof. We have defined the evolute as the envelope of a one-parameter family of normals of the curve L. This family is specified by equation (16.45). Thus, the function $F(X, Y, t)$ defining the family is determined by the relation

$$F(X, Y, t) = \varphi'(X - \varphi) + \psi'(Y - \psi), \tag{16.47}$$

the quantity t serving as a parameter.

Let us now apply the conclusions derived in 16.1 concerning the existence of the envelope to solve the problem of the existence of an evolute (i.e. the envelope of family (16.45)). We must verify whether the conditions of the lemma from 16.1.3 and the conditions of Theorem 16.1 have been fulfilled. We shall first verify the conditions of the lemma. Evidently, with the requirements we have placed upon the functions φ and ψ, the functions $F(X, Y, t)$ and $F'_t(X, Y, t)$ are differentiable. Let us make sure that the Jacobian $\frac{D(F, F'_t)}{D(X, Y)}$ is different from zero. Making use of expression (16.40) for the curvature k of the curve L, it is possible to represent this Jacobian in the form

$$\frac{D(F, F'_t)}{D(X, Y)} = k[\varphi'^2+\psi'^2]^{3/2}.$$

By the hypothesis, the curvature k is nonzero. Besides, since all the points of the curve L are regular, we have $\varphi'^2 + \psi'^2 \neq 0$. Thus we see that the indicated Jacobian is different from zero and, consequently, all the conditions of the lemma are fulfilled.

Let us now verify the conditions of Theorem 16.1. It is evident that with the requirements placed upon the functions φ and ψ, the derivatives F'_X, F'_Y, F''_{tX}, F''_{tY}, F''_{tt} are continuous. It remains to verify the validity of the relation $F''_{tt} \neq 0$ for all the values of t belonging to the interval (0, 1) and for the characteristic points on the normals to the curve L.

* Under the conditions imposed on the functions φ and ψ the curvature k is a differentiable function of the parameter t.

Since

$$F''_{tt} = \varphi'''(X - \varphi) + \psi'''(Y - \psi) - 3(\varphi'\varphi'' + \psi'\psi'') \tag{16.48}$$

and in accordance with 16.1.2 and formulas (16.13), for the case under consideration, the characteristic points are determined by the relations

$$\left.\begin{aligned} &\varphi'(X - \varphi) + \psi'(Y - \psi) = 0, \\ &\varphi''(X - \varphi) + \psi''(Y - \psi) - (\varphi'^2 + \psi'^2) = 0, \end{aligned}\right\} \tag{16.49}$$

it follows that by virtue of (16.48) and (16.49) the value of the derivative F''_{tt} at the characteristic points is equal to

$$F''_{tt} = \frac{\varphi'\psi''' - \varphi'''\psi'}{\varphi'\psi'' - \varphi''\psi'}(\varphi'^2 + \psi'^2) - 3(\varphi'\varphi'' + \psi'\psi''). \tag{16.50}$$

Let us turn now to expression (16.40) for the curvature k of the curve L. Taking into consideration the notation of (16.41), we obtain the following expression from formula (16.40) by means of differentiation:

$$k' = \frac{\varphi'\psi'' - \varphi''\psi'}{[\varphi'^2 + \psi'^2]^{5/2}}\left\{\frac{\varphi'\psi''' - \varphi'''\psi'}{\varphi'\psi'' - \varphi''\psi'}(\varphi'^2 + \psi'^2) - 3(\varphi'\varphi'' + \psi'\psi'')\right\}.$$

With the aid of this relation and expression (16.40) for the curvature, we shall impart to expression (16.50) the following form:

$$F''_{tt} = \frac{k'}{k}(\varphi'^2 + \psi'^2).$$

Since by the hypothesis k and k' are nonzero and, besides, $\varphi'^2 + \psi'^2 \neq 0$, it follows from the last expression for F''_{tt} that $F''_{tt} \neq 0$ for all the values of t from the interval $(0, 1)$ and for the characteristic points on the normals to L. Thus the conditions of Theorem 16.1 are also fulfilled.

Let us now find the parametric equations of the evolute. Since under the conditions we have formulated the evolute is the locus of the characteristic points of the family of normals, the coordinates X and Y of the points of the evolute are determined by relations (16.49). Having found X and Y from these relations, we shall obtain parametric equations (16.46). The theorem is proved.

Remark. The notions of the radius of curvature and the centre of curvature are often employed in geometry. The *radius R of curvature* of the curve L is the quantity $1/k$, where k is the curvature of L, and the *centre of curvature* is the point of the normal to the curve which is at a distance R from the given point of the curve L in the direction of concavity of L.

Note that the radius of curvature of a curve is equal to the radius of an osculating circle, and the centre of curvature coincides with the centre of the osculating circle.

Let us verify that the *evolute is the locus of the centres of curvature of a curve.* Indeed, we obtain from (16.46)

$$(X-\varphi)^2+(Y-\psi)^2=\frac{1}{k^2}=R^2,$$

that is, the point of the evolute having the coordinates (X, Y) is at the distance R from the point of the curve L with the coordi-

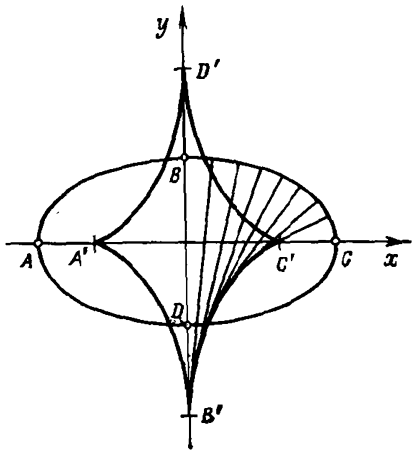


Fig. 16.16

nates (φ, ψ). The geometrical meaning of the evolute (see Fig. 16.15) implies that point (X, Y) is located on the normal in the direction of concavity of the curve L.

Example. Find the equation of the evolute of the ellipse $ABCD$ (Fig. 16.16) specified by the parametric equations

$$x=\varphi(t)=a\cos t,\quad y=\psi(t)=b\sin t$$

Since $\varphi'=-a\sin t$, $\psi'=b\cos t$, $\varphi''=-a\cos t$, $\psi''=-b\sin t$, it follows that $\varphi'^2+\psi'^2=a^2\sin^2 t+b^2\cos^2 t$, and $\varphi'\psi''-\varphi''\psi'=ab$. Therefore, in accordance with (16.46), the parametric equations of the evolute of the ellipse have the form

$$X=\frac{a^2-b^2}{a}\cos^3 t,\quad Y=\frac{b^2-a^2}{b}\sin^3 t.$$

Thus we see that the *evolute of an ellipse* is the so-called *elongated astroid* (see Fig. 16.16).

Note that at the points A, B, C, D of the ellipse (i.e. at its vertices) the derivative of the curvature is equal to zero. Therefore, the conditions of Theorem 16.4 are not fulfilled at these points. At the respective points A', B', C', D' the evolute of the ellipse has peculiarities, the so-called cuspidal points (cusps).

APPENDIX

FURTHER DEVELOPMENT OF THE THEORY OF REAL NUMBERS

To introduce real numbers in Chapter 2, we have made use of a set of infinite decimal fractions. Having defined the rules of comparison, addition and multiplication for the set of these fractions, we established that the members of the set possess 13 *basic properties* (enumerated in 2.1.1 for rational numbers). Although the method of introducing real numbers described there has considerable heuristic and methodical merits, it is not the only one possible. We could have introduced real numbers with the aid of infinite binary fractions, with the aid of the so-called Dedekind cuts in the field of rational numbers*, with the aid of sequences of rational numbers** or by some other methods. To elucidate the interconnection between various methods of introduction of real numbers, we shall employ some new notions and establish an important property of the set of the real numbers we have studied.

1. Completeness of a set of real numbers. Two sets for whose members the rules of comparison, addition and multiplication are defined are said to be *isomorphic to each other with respect to these rules* if a one-to-one correspondence*** can be established between the members of these sets so that if the members a and b of the first set are associated with the members a' and b' of the second set, then (1) the members a' and b' are connected by the same sign ($>$, $<$ or $=$) as the members a and b; (2) the member $a' + b'$ corresponds to the member $a + b$; (3) the member $a' \cdot b'$ corresponds to the member $a \cdot b$.

We could use analogy to speak of some other rules, besides comparison, addition and multiplication, characterizing the relations between the members, and to introduce the notion of the sets isomorphic to each other with respect to the indicated rules.

We can cite as an example of two sets isomorphic to each other with respect to the rules of comparison, addition and multiplication the set of rational numbers introduced as a ratio of integers obeying

* The method of introducing real numbers with the aid of cuts belongs to the German mathematician Richard Dedekind (1831-1916). You can find the exposition of this method in the first chapter of *A Treatise on Advanced Calculus* by Philip Franklin, for instance.

** This method of introduction of real numbers belongs to Georg Cantor.

*** For a one-to-one correspondence of the members of two sets, see the first footnote on p. 85.

the rules of comparison, addition and multiplication, indicated in the footnotes on p. 36, and the set of rational numbers written as infinite decimal fractions obeying the ordinary rules of comparison, addition and multiplication of real numbers.

Let us consider two sets more attentively, the set of all rational numbers and the set of all real numbers. The rules of comparison, addition and multiplication are defined for each of these sets and the 13 basic properties are valid for them. It is clear, however, that the set of real numbers is "broader" than the set of all rational numbers since *as a whole the set of all real numbers is not isomorphic to the set of all rational numbers as concerns the rules of comparison, addition and multiplication* but we can isolate a part of the set of real numbers which is isomorphic to the set of all rational numbers with respect to these rules.*

A natural question arises whether it is possible to construct a "broader" set of objects for the set of real numbers as well which would possess the following properties: (1) the rules of comparison, addition and multiplication are defined and the thirteen basic properties are valid for this "broader" set, (2) as a whole, the "broader" set is not isomorphic to the set of all real numbers with respect to the indicated rules, (3) we can isolate from the "broader" set a part which is isomorphic to the set of all real numbers as concerns the indicated rules.

We shall prove that such a "broader" set *does not exist*, that is, as it is customary to say in mathematics, *the set of all real numbers is* **complete** *with respect to the rules of comparison, addition and multiplication and the 13 basic properties.*

In general, *an arbitrary set of objects for which certain rules are defined and certain properties are valid is said to be complete with respect to these rules and properties if it is impossible to construct a "broader" set of objects such that* (1) *the same rules would be defined and the same properties would be valid for this "broader" set*, (2) *this "broader" set would not be isomorphic, as a whole, to the given set as concerns the indicated rules*, (3) *there would be a part of the "broader" set isomorphic to the given set with respect to the indicated rules.*

It can be stated that the set of all rational numbers *is not complete with respect to the rules of comparison, addition and multiplication and to the thirteen basic properties*, since there is a "broader" set (the set of real numbers) satisfying requirements (1), (2) and (3) from the definition just formulated.

Let us prove now that *the set of all real numbers is complete with respect to the rules of comparison, addition and multiplication and to the thirteen basic properties.*

* It follows from the fact that a one-to-one correspondence cannot be established between the set of all rational numbers and the set of all real numbers (see 3.4.6).

Let us assume the contrary, that is, let us suppose that there exists a "broader" set of objects $\{x'\}$ such that (1) the rules of comparison, addition and multiplication are defined and the thirteen basic properties are valid for the members of the set $\{x'\}$, (2) as a whole, the set $\{x'\}$ is not isomorphic to the set $\{x\}$ of all real numbers with respect to the indicated rules, (3) there is a part of the set $\{x'\}$ (we shall designate it as $\{\bar{x}'\}$) which is isomorphic to the set $\{x\}$ of all real numbers with respect to the indicated rules.

Note, first of all, that the set $\{x'\}$ contains a *single pair* of members $0'$ and $1'$ playing a particular role of zero and unity*. We can then state that the members $0'$ and $1'$ enter into the set $\{\bar{x}'\}$ and are in a one-to-one correspondence with the real numbers 0 and 1**. Assume that α' is some member of the set $\{x'\}$ which *does not belong* to the set $\{\bar{x}'\}$.

By virtue of the comparison rule, we can divide all the members $\bar{x}'$ of the set $\{\bar{x}'\}$ into *two classes, the upper class and the lower class*, placing into the upper class all the members $\bar{x}'$ satisfying the inequality $\bar{x}' > \alpha'$, and into the lower class all the members $\bar{x}'$ satisfying the inequality $\bar{x}' < \alpha'$. Both these classes are nonempty. Let us prove, for instance, that the upper class is nonempty. Repeating the member $1'$ as a summand a sufficient number of times, we shall obtain, by virtue of property 13°, a member n' of the set $\{\bar{x}'\}$ satisfying the inequality $n' > \alpha'$, that is, belonging to the upper class. It follows from property 1° that *every member of the lower class is smaller than any member of the upper class.* Because of the isomorphism of the set $\{\bar{x}'\}$ and the set $\{x\}$ of all the real numbers, we can assert that the set of all real numbers can also be divided into two classes with every number from the lower class being smaller than any number from the upper class. But this means that the *lower class of real numbers is bounded from above and possesses* (by virtue of Theorem 2.1) *the least upper bound* (*the supremum*) m while the upper class possesses the greatest lower bound (the infimum) M. It follows from the definition of the supremum and the infimum that both bounds m and M

* If there were two members $0'_1$ and $0'_2$ playing a particular role of zero, then, by virtue of the properties of the sum, we should have $0'_1 = 0'_1 + 0'_2 = 0'_2 + 0'_1 = 0'_2$, i.e. $0'_1 = 0'_2$. We can prove by analogy the uniqueness of the member $1'$ playing a particular role of unity.

** Let us prove, for instance, that the member $0'$ enters into the set $\{\bar{x}'\}$ and is in a one-to-one correspondence with the real number 0. Let us suppose that the real number 0 is associated with some member θ' of the set $\{\bar{x}'\}$ and a' is any member of this set corresponding to the real number a. By virtue of the isomorphism with respect to addition, the member $a' + \theta'$ is associated with the real number $a + 0 = a$, i.e. the member $a' + \theta'$ coincides with a', but this means that θ' is the (unique!) zero member, i.e. $\theta' \equiv 0'$. The argument for the unity member is the same.

are contained between real numbers arbitrarily close to each other and, therefore, $m = M$. Since the number $m = M$ is one of the real numbers, it belongs to one of the classes, that is, *there exists either the smallest member in the upper class or the greatest member in the lower class*. Let us prove the absurdity of the two assertions. Suppose (for definiteness) that there exists the smallest member in the upper class of real numbers. Then there also exists the smallest member m' in the upper class corresponding to the partitioning of the set $\{\bar{x}'\}$. By the definition of the upper class $m' > \alpha'$. In accordance with the properties of the sum, there is a difference $m' - \alpha'$, with $m' - \alpha' > 0'$ by virtue of these properties. But then, by Property 9°, there is an inverse of the member $m' - \alpha'$ which, by virtue of the properties of the product, is equal to the quotient $\frac{1'}{m'-\alpha'}$. In accordance with Property 13°, the member $1'$ can be repeated as a summand a sufficient number of times for the resulting "integral" member n' to belong to $\{\bar{x}'\}$ and satisfy the inequality $n' > \frac{1'}{m'-\alpha'}$. By virtue of the properties of the product and the sum, we obtain from the last inequality*

$$m' - \frac{1'}{n'} > \alpha'. \tag{A.1}$$

Since the members m', $1'$ and n' belong to the set $\{\bar{x}'\}$, the member $\left(m' - \frac{1'}{n'}\right)$ also belongs to that set and, evidently, satisfies the inequality $m' - \frac{1'}{n'} < m'$. But then inequality (A.1) signifies that *there is a member in the upper class which is smaller than m', i.e. m' is not the smallest member*. The contradiction obtained proves the completeness of the set of real numbers.

2. Axiomatic introduction of the set of real numbers. The axiomatic method of introducing real numbers serves as a logical completion of the conception of these numbers. This method consists in the following.

*The set of real numbers is introduced as a totality of objects of any kind satisfying 17 axioms which are the rules of comparison, addition and multiplication***, *the thirteen basic properties, and the axiom on the completeness with respect to the indicated rules and properties.*

These 17 axioms are usually called the *axioms of a real number*. A specific realization of the totality of objects satisfying the seventeen axioms of a real number is the set of infinite decimal fractions

* These properties ensure the applicability of all the rules of algebra.
** Only the fact of the existence of the rules of comparison, addition and multiplication is postulated, the specific form of the rules is not indicated.

we have discussed in Chap. 2. Other realizations of this totality of objects are also possible*. The connection between these realizations becomes quite clear from the following remarkable assertion.

Any realization $\{x'\}$ of the totality of objects satisfying the seventeen axioms of a real number is isomorphic, with respect to the rules of comparison, addition and multiplication, to the set $\{x\}$ of infinite decimal fractions discussed above.

Proof. For the sake of convenience, we shall divide the proof into separate stages.

1°. We shall first note that the axioms ensure the possession by the set $\{x'\}$ of the members $0'$ and $1'$ playing a particular role of zero and unity. In accordance with the Archimedean axiom $1' > 0'$**. Let us isolate from the set $\{x'\}$ a totality of "rational objects". For that purpose, we make note of the fact that any rational number can be obtained from the numbers 0 and 1 by means of the operations of addition, subtraction and division. Indeed, repeating the number 1 as a summand a requisite number of times, we can obtain any positive integer n; subtracting the number 1 from the number 0 a requisite number of times, we shall obtain any negative integer; a division of one integer into another can yield any rational number. Since, in accordance with the axioms, the operations of addition, subtraction and division are defined for the set $\{x'\}$, these operations will help us to obtain all "rational objects" from $0'$ and $1'$. We shall use the same notation for these objects as for rational numbers but will supply them with primes.

Let us prove that the totality of "rational objects" of the set $\{x'\}$ we have constructed is isomorphic with respect to the rules of comparison, addition and multiplication of the totality of rational numbers of the set $\{x\}$. In fact let us associate the rational number $\frac{m}{n}$ with the "rational object" $\frac{m'}{n'}$. It follows from the method used for constructing "rational objects" that the sum and the product of rational numbers $\frac{m}{n}$ and $\frac{p}{q}$ correspond to the sum and the product of the "rational objects" $\frac{m'}{n'}$ and $\frac{p'}{q'}$. It remains to verify that $\frac{m'}{n'}$ and $\frac{p'}{q'}$ are connected by the same sign as $\frac{m}{n}$ and $\frac{p}{q}$. When we construct "rational ob-

* We have indicated in the introduction to the present appendix that real numbers can be introduced by various methods (by means of infinite binary fractions, by the so-called Dedekind cuts and by other methods).

** Indeed, if the opposite inequality $1' \leqslant 0'$ were valid, then, by virtue of the axioms, we would obtain from it $1' + 1' + \ldots + 1' \leqslant 0' + 0' + \ldots + 0' = 0'$ (whatever the number of times $1'$ is repeated as a summand), and this contradicts the Archimedean axiom.

jects", the rule of comparison of any "rational objects" by multiplying them by "integral objects" reduces to a comparison of "integral objects", therefore, it is sufficient to verify that $m' > n'$ at $m > n$ for any "integral objects" m' and n'. To verify this fact, it is sufficient, by virtue of the axioms, to prove that $(n+1)' > n'$. The last inequality follows from the fact that $1' > 0'$ and, hence,

$$(n+1)' = n' + 1' > n' + 0' = n'.$$

2°. Assume now that a' is any object of the set $\{x'\}$. Let us show that *this object can be associated with a definite "infinite decimal fraction"*. We shall assume for definiteness that $a' > 0'$. By virtue of the Archimedean axiom, among the "integral objects", strictly smaller than a', there is the greatest object which will be denoted by a_0'; among the "rational objects"

$$a_0', 0'; \quad a_0', 1'; \quad \ldots; \quad a_0', 9',$$

which are strictly smaller than a', there is the greatest object which will be denoted by a_0', a_1', etc.

Thus we shall associate any object with an infinite totality of "rational objects"

$$a_0'; \; a_0', \; a_1'; \; \ldots; \; a_0', \; a_1' \; \ldots \; a_n'; \; \ldots, \tag{A.2}$$

or, what is the same, an "infinite decimal fraction"

$$a_0', \; a_1'a_2' \; \ldots \; a_n' \; \ldots \tag{A.3}$$

The same argument is valid for $a' < 0'$, but in this case all the objects (A.2) and the "infinite decimal fraction" (A.3) will have a minus sign.

It is evident from the construction of the totality of objects (A.2) that the inequalities

$$a_0', \; a_1'a_2' \; \ldots \; a_n' < a' \leqslant a_0', \; a_1'a_2' \; \ldots \; a_n' + \frac{1'}{(10^n)'} \tag{A.4}$$

are valid for any number n, that is, any object a' is contained between two "rational objects" whose difference $\frac{1'}{(10^n)'}$ can be made smaller than any preassigned "positive" object *.

3°. *Let us prove now that if two objects a' and b' can be contained between two "rational objects" α' and β' $(\beta' > \alpha')$, whose difference $\beta' - \alpha'$ can be made smaller than any preassigned "positive" object, then $a' = b'$.* Let us assume that $a' \neq b'$, say, $a' < b'$. Then $\alpha' \leqslant$

* Indeed, by virtue of the axioms, for any object $\varepsilon' > 0'$ there is an inverse $\frac{1'}{\varepsilon'}$ and there is an "integral object" n' for the latter such that $n' > \frac{1'}{\varepsilon'}$, so that $\frac{1'}{(10^n)'} < \frac{1'}{n'} < \varepsilon'$.

$\leqslant a' < b' \leqslant \beta'$. By virtue of the axioms, we shall obtain from these inequalities

$$0' < b' - a' \leqslant \beta' - \alpha'. \tag{A.5}$$

But then the object $b' - a'$ will have an inverse $\frac{1'}{b' - a'}$, and the latter will have an "integral object" n' such that

$$n' > \frac{1'}{b' - a'},$$

so that

$$\frac{1'}{n'} < b' - a'.$$

Comparing the last inequality with (A.5), we obtain

$$\frac{1'}{n'} < \beta' - \alpha',$$

which contradicts the fact that the difference $\beta' - \alpha'$ can be made smaller than any preassigned "positive" object.

4°. Let us verify that *two unequal objects of the set* $\{x'\}$ *are associated with various "infinite decimal fractions"*. In fact, let us suppose that two objects from $\{x'\}$ are associated with one and the same "infinite decimal fraction" (say, (A.3)). Then, in accordance with inequalities (A.4), these two objects can be contained between "rational objects" whose difference can be made smaller than any preassigned "positive" object. On the basis of 3° the objects in question are equal. The assertion we have proved justifies the representation of any object of the set $\{x'\}$ by an "infinite decimal fraction".

5°. In accordance with the first three axioms, the rules of comparison, addition and multiplication are defined for the objects of the set $\{x'\}$. Let us prove that *if all the objects of the set* $\{x'\}$ *are represented by "infinite decimal fractions", then the rule of comparison and the definition of the sum and the product for these "fractions" are formulated in the same way as for ordinary infinite decimal fractions discussed above.*

Assume that a' and b' are any two objects of the set $\{x'\}$ and let these objects be associated with the "infinite decimal fractions"*

$$a'_0, a'_1 \ldots a'_n \ldots \text{ and } b'_0, b'_1 \ldots b'_n \ldots \tag{A.6}$$

Let us first establish the rule of comparison of objects a' and b' represented as "infinite decimal fractions" (A.6). It is sufficient to prove the following two assertions: (1) if the fractions (A.6) coincide,

* When establishing the rule of comparison, it is possible to restrict the study to "positive" objects a' and b', since the general case can be reduced to this case with the aid of the rule of signs.

that is, if

$$a'_0 = b'_0,\ a'_1 = b'_1,\ \ldots,\ a'_n = b'_n,\ \ldots,$$

then the objects a' and b' are equal; (2) if there is a number n such that the relations

$$a'_0 = b'_0,\ a'_1 = b'_1,\ \ldots,\ a'_{n-1} = b'_{n-1},\ a'_n < b'_n \tag{A.7}$$

hold true, then the objects a' and b' are related by the inequality $a' < b'$.

We have proved assertion (1) in 4°. Let us prove assertion (2). The last relation (A.7) can be rewritten as

$$a'_n + 1' \leqslant b'_n.$$

Using this relation and the remaining relations (A.7) and writing the right-hand inequality (A.4) for a' and the left-hand inequality (A.4) for b', we obtain

$$a' \leqslant a'_0,\ a'_1 \ldots a'_{n-1}a'_n + \frac{1'}{(10^n)'} = a'_0,\ a'_1 \ldots a'_{n-1}(a'_n + 1') \leqslant$$

$$\leqslant a'_0,\ a'_1 \ldots a'_{n-1}b'_n < b',\ \text{i.e. } a' < b'.$$

Let us prove that the same definitions of the sum and the product are valid for the objects a' and b' as for ordinary real numbers. We shall restrict our discussion to the case of the sum. Suppose α'_1, α'_2, β'_1 and β'_2 are various "rational objects" satisfying the inequalities

$$\alpha'_1 \leqslant a' \leqslant \alpha'_2,\ \beta'_1 \leqslant b' \leqslant \beta'_2. \tag{A.8}$$

Then the sum $a' + b'$ of the objects a' and b' is the only object satisfying the inequalities

$$\alpha'_1 + \beta'_1 \leqslant a' + b' \leqslant \alpha'_2 + \beta'_2. \tag{A.9}$$

Indeed, the axioms yield the possibility of termwise addition of the inequalities, whence it follows that the sum $a' + b'$ complies with inequalities (A.9). Besides, the sum is the only object satisfying inequalities (A.9) since each of the differences

$$\alpha'_2 - \alpha'_1 \text{ and } \beta'_2 - \beta'_1$$

and, consequently, the difference

$$(\alpha'_2 + \beta'_2) - (\alpha'_1 + \beta'_1)$$

can be made smaller than any preassigned "positive" object (by virtue of 2°). The case of the product can be treated analogously.

6°. Let us finally prove that the set $\{x'\}$ is isomorphic to the set of real numbers representable by "infinite decimal fractions". Let us assume that the set $\{x'\}$ is not isomorphic to $\{x\}$. We shall associ-

ate the object a', representable by an "infinite decimal fraction" $a'_0a'_1a'_2 \ldots$, with a real number $a = a_0, a_1, a_2 \ldots$. Suppose that two objects a' and b' are associated with the real numbers a and b.

It follows from the results of A.5° that (1) a' and b' are connected by the same sign as the numbers a and b; (2) the sum $a' + b'$ is associated with the sum $a + b$; (3) the product $a' \cdot b'$ is associated with the product $a \cdot b$*.

Since by the assumption $\{x'\}$ is not isomorphic to $\{x\}$, it follows that with the indicated relationship not every real number is associated with a certain member of the set $\{x'\}$. This means that the set $\{x'\}$, not being isomorphic to the entire set $\{x\}$, is isomorphic to a part of the set $\{x\}$, which contradicts the completeness of the set $\{x'\}$.

The contradiction obtained completes the proof of the assertion.

3. Concluding remarks. We want to point out in conclusion that the axiomatic method and the notion of isomorphic (with respect to various rules) totalities of objects are widely used in various branches of modern mathematics and physics (in geometric constructions, in the probability theory, classical mechanics, statistical physics, quantum mechanics** and other fields).

For instance, in geometry the set of points of a straight line is introduced as a totality of objects satisfying certain axioms, the principal role being played by the axiom on the completeness of that totality with respect to the rest of the axioms. These axioms make it possible to establish a one-to-one correspondence between the set of all points of the straight line and the set of all real numbers. This correspondence enables us to represent real numbers by points on the straight line (number axis), which fact is widely used in the course of analysis and as a means of illustrating the argument.

---

* Since the comparison, addition and multiplication of the objects a' and b' and of the real numbers a and b are defined by the same rules.

** Quantum mechanics originated as two apparently different theories: "Matrix mechanics" initiated by Heisenberg and "Wave mechanics" initiated by Schrödinger. It was proved later on that these two theories used two specific realizations, isomorphic to each other, of one common totality of objects introduced axiomatically and known as Hilbert's abstract space (see Part 2 of the present course).

NAME INDEX

Abel, Niels Henrik, 363
Archimedes, 33

Bolzano, Bernhard, 80
Bonnet, P. O., 349
Borel, Emile, 337
Budak, B. M., 6
Bunyakovsky, V. Ya., 359

Cauchy, Augustin Louis, 82
Cesàro, Ernesto, 468
Chebyshev, P. L., 229
Cohen, Paul Joseph, 86
Coulomb, Charles Augustin, 18

D'Alembert, Jean Le Rond, 434
Darboux, Gaston, 322
De Moivre, Abraham, 200
Dirichlet, Peter Gustav Lejeune, 95

Efimov, N. V., 5
Euclid, 205
Euler, Leonhard, 23

Fomin, S. V., 6
Franklin, Philip, 5, 621
Fresnel, Augustin, 189

Gaisaryan, S. S., 556
Goursat, Edouard Jean Baptiste, 5

Hadamard, Jacques, 591
Hardy, Godfrey Harold, 471
Heine, Eduard, 337
Hölder, Otto, 358

Jacobi, Karl Gustav, 571
Jordan, Camille, 380

Kudryavtsev, L. D., 6

Lagrange, Joseph Louis, 257
Laplace, Pierre Simon, 593
Lebesgue, Henri, 337
Legendre, Adrien Marie, 239
Leibniz, Gottfried Wilhelm von, 32
L'Hospital, Gullaume François Antoine de 264
Liouville, Joseph, 239

Maclaurin, Colin, 275
Minkowski, Hermann, 359
Newton, Isaac, 17

Ostrogradsky, M. V., 221

Peano, Guiseppe, 273
Poisson, Siméon Denis, 469
Poussin, De La Vallée, 5

Raabe, Josef Ludwig, 440
Rouché, Eugène, 270
Riemann, Bernhard, 325
Rolle, Michel, 256

Schlömilch, Oskar Xaver, 270
Simpson, Thomas, 420
Stolz, Otto, 89
Sveshnikov, A. G., 6
Sylvester, James Joseph, 527

Taylor, Brook, 269
Tikhonov, A. N., 5

Wallis, John, 463
Weierstrass, Karl, 80

Abel's transformation, 363
Abel-Dirichlet test, 454
Abel-Poisson summation method, 469
Absolute convergence of an infinite product, 462
Absolute value of a real number, 41
Absolutely convergent series, 443
Algebraic polynomial, 201
Algorithm, 34
  Euclidean, 205
Alternating series, 453
Analytic method of representing a function, 19, 95
Antiderivative, 184
Archimedes' axiom, 37
Area
  of a curvilinear sector, 383
  of a curvilinear trapezoid, 382
  lower, 380
  of a surface of revolution, 389
  upper, 380
Argument, 18, 94
  increment of, 150
  intermediate, 19
Asymptote of a graph, 313
  inclined, 314
  vertical, 313
Astroid, 389

Basic concepts, 18
Bolzano-Weierstrass theorem, 80, 481
Borel-Heine lemma, 337
Bound
  of a function (least upper, greatest lower), 251
  of a number set (upper, lower), 43
Boundary point of a set, 477
Bounded
  domain, 477
  function, 247
  sequence, 57, 481
  set, 84
Boundedness of a local function, 248
Bunyakovsky inequality (for sums), 359
Bunyakovsky-Cauchy inequality (for integrals), 362

Cauchy
  form of remainder, 273
  formula, 263
  test for convergence of a sequence, 83, 479
  test for convergence of a series, 435
  test for existence of a limiting value, 244, 483
  theorem, 448
Cauchy-Bunyakovsky inequality for integrals, 362
Cauchy-Maclaurin test, 437
Catenary, 379
Centre of curvature, 619
Cesàro method of summation, 468
Change of variable, 353
Charasteristic, 603
Characteristic point, 598
Chord method, 401, 409
Closed
  interval, 53
  set, 477
Combined method of seeking roots, 411
Comparison tests for series, 430
Complex numbers, 197

Completeness of a set of real numbers, 622
Composite function, 24, 132, 489
  continuity of, 105, 489
  differentiation of, 24, 169, 499
Conditionally convergent
  product, 461
  series, 441
Constant quantity, 17
Contact of curves, 604
Continuity of a function, 22, 103, 485
  difference form, 151, 486
  at a point, 103, 243, 485
  on the right (on the left), 104
  on a set, 104, 486
  on a line, 487
  uniform, 334, 491
Continuity of a composite function, 489
Convergent
  product, 458
  sequence, 63, 479
  series, 424
Convex function, 536
  strictly, 540
  strongly, 542
Convex set, 536
Coordinate
  plane, 472
  space, 473, 475
Coulomb law, 18
Cubable body, 387
Curvature of a curve, 610
Curve
  arc length of, 370
  simple, 364
Curvilinear
  sector, 383
  trapezoid, 30, 382
Cycloid, 183, 379

D'Alembert's test, 434
Darboux
  integrals, 330
  lemma, 331
  theorem, 322
De Moivre formula, 200
Decrease of a function at a point, 254
Definite integral, 30, 325
Derivative, 20, 151
  directional, 505
  of higher order, 177
  partial, 491
  right-hand, left-hand, 154
  of vector function, 154
Diameter of a set, 491
Difference of real numbers, 50
Differentiable function, 156, 493
  n times, 177, 509
Differential, 157, 499
  of an arc, 378
  of higher orders, 177, 511
  invariance of form, 173, 503
Differentiation, 20, 157
  of a composite function, 24, 169, 499
  of a sum, difference, product, quotient, 24, 159
Dirichlet function 95
Dirichlet-Abel test, 454
Discontinuity of a function
  of the first kind, 138
  removable, 137
  of the second kind, 138
Discriminant
  curve, 600
  surface, 603
Divergent series, 424
  summation methods, 468
Domain, 477
  bounded, 477
  closed, 477
  of definition of a function, 94, 479

Elementary function, 110, 137
Ellipse, 379
Elliptic integral, 238
Envelope
  of a family of curves, 599
  of a family of surfaces, 603

Equivalent infinitesimal functions, 101
Euclidean
  algorithm, 205
  plane, 473
  space, 473, 475
Euler's
  substitution, 230
  theorem, 502
Evolute, 617
Exponential function, 112
Extremum
  end-point, 322
  local, 255

"Fork" method, 399
Formula of trapezoids, 417
Fresnel integral, 189
Function(s), 17, 94, 479
  bound of (least upper, greatest lower), 251
  bounded, 246
  composite, 24, 105, 132, 489
  condition of continuity, difference form, 151
  continuity, 22, 103, 485
  convex, 536, 540
  decrease of (at a point), 254
  domain of definition, 94
  differentiable, 156, 481
  Dirichlet, 95
  elementary, 137
  equivalent infinitesimal, 101
  exponential, 112
  graph of, 474
  homogeneous, 502
  hyperbolic, 117
  implicit, 559
  increase of (at a point), 254
  increment of, 150
  infimum of, 251
  infinitely large at a point, 101
  infinitesimal, 100
  integrability of, 325
  inverse, 106, 569
  limiting value of, 96, 241, 479, 482
  local, 248
  logarithmic, 116
  methods of representation of, 19, 95
  monotonic, 106
  oscillation of, 334, 491
  particular value of, 94
  piecewise continuous, 140
  point of discontinuity of, 104
  power, 118
  range of, 94
  strictly convex, 540
  strongly convex, 542
  supremum of, 251
  trigonometric, 121
  vector, 154
Fundamental sequence, 82, 480

Generalized harmonic series, 433
Geometric progression, 425
Gradient, 505
Gradient method of seeking extremum, 535
Graph of a function, 26, 474
  direction of convexity, 304
  point of inflection, 306
Graphical method of representing a function, 19, 96

Hadamard inequality, 591
Half-open (half-closed) interval, 53
Half-line, 53
Harmonic oscillation, 19
Harmonic series, 428
  generalized, 433
Heine-Borel lemma, 337
Hölder's inequality
  for integrals, 360
  for sums, 358
Homeomorphic mapping, 577
Homogeneous function, 502
Hyperbolic function, 117

Implicit function, 559
Improper rational fraction, 209
Inclined asymptote, 314

Increase of a function at a point, 254
Increment of an argument, 150
Increment of a function, 150
  total, 486
  partial, 486
Indefinite integral, 27, 184
Infimum of a function, 251
Infimum of a set, 44
Infinite
  decimal fraction, 39
  interval, 53
  product, 457
Infinitely large
  function at a point, 101
  sequence, 59
Infinitesimal
  function, 100
  sequence, 59
Integrability
  condition of, 332
  of a function, 325
Integral
  definite, 30, 325
  elliptic, 238
  indefinite, 27, 184
Integral cosine, sine, 189
Integral logarithm, 189
Integral sum, 324
Integration
  of binomial differentials, 228
  by a change of variable, 190, 353
  of linear-fractional irrational functions, 227
  by parts, 193, 354
  of a rational fraction, 218
Interior point
  of a closed interval, 53
  of a set, 477
Intermediate argument, 19
Interval, 53
Invariance of form of a differential, 173, 503
Inverse function, 106, 569
  trigonometric, 124
Involute, 617
Irrational numbers, 40
Isometric sets, 621
Iteration, 90
Iteration method, 402
Iterative sequence, 403

Jacobi's determinant, 571

Lagrange's
  form of remainder, 273
  formula of finite increments, 257
  function, 587
  method of multipliers, 586
Laplacian operator, 593
Law of motion, 17
Leibniz formula, 179
Leibniz series, 453
Leibniz-Newton formula, 32, 352
Length of arc of a curve, 370
Level surface, 506
L'Hospital's rule, 264
Limit
  of a sequence, 63, 479
  inferior of a sequence, 78
  superior of a sequence, 78
Limit point of a sequence, 75
Limiting value of a function, 96, 241, 479, 482
  necessary and sufficient condition for existence, 483
Local extremum
  of functions of several variables, 524
  necessary and sufficient condition for, 524
Local function
  boundedness of, 248
  extremum of (maximum, minimum), 525
Logarithm
  integral, 189
  natural, 117
Logarithmic function, 116
Lower sum, 328

Maclaurin's formula, 275
Maclaurin-Cauchy test, 437

Mapping, one-to-one, 576
homeomorphic, 577
Mean value
first formula, 347, 348
second formula, 349
Method of interpolation, 96
Method of parabolas (Simpson's rule), 419, 420
Method of rectangles, 414
Method of undetermined coefficients, 216
Methods of representing a function
analytical, 19, 95
graphical, 19, 96
tabular, 19, 96
Metric space, 475
Minkowski's inequality
for integrals, 362
for sums, 359
Monotonic
function, 106
sequence, 68

Natural logarithm, 117
Natural parameter on a curve, 371
Neighbourhood of a point, 53, 477, 478
Newton's
binomial formula, 21
method of tangents, 400, 406
polynomial formula, 517
Newton-Leibniz formula, 32, 352
Normal to a surface, 497
Number *e*, 23, 73, 280
Number
line, 53
sequence, 57
series, 424

One-parameter family of curves, 598
One-sided limiting value, 97
One-to-one
mapping, 576
correspondence, 85
Open
half-line, 53
interval, 53
Order of contact of curves, 605
Oscillation of a function, 334, 491
Osculating circle, 609
Ostrogradsky method, 221

Partial derivative, 491
of higher order, 507
Partial sum of a series, 424
Particular value of a function, 94, 479
Peano form of a remainder, 274
Piecewise continuous function, 140
Point of discontinuity of a function, 104
classification of, 137
Poisson integral, 189
Poisson-Abel summation method, 469
Power function, 118
Product
infinite, 457
of real numbers, 49
Proper rational fraction, 209

Quadratic form, 514
alternating, 527
Quantity (variable, constant), 17
Quotient of real numbers, 51

Raabe's test, 440
Radius of curvature, 619
Range of a function, 94
Rational fraction, 209
improper, 209
proper, 209
Rational numbers, 35
Real number(s), 35, 40, 621
absolute value of, 41
difference of, 50
irrational, 40
rational, 35
rule of comparison, 41
product of, 49
quotient of, 51
sum of, 47
Rectangles, method of, 414

Rectangular neighbourhood of a point, 477
Rectifiable curve, 370
Reduced polynomial, 203
Regular (nonsingular) point, 568, 595
Remainder
  in the form of Cauchy, 273
  in the form of Lagrange, 273
  in the form of Schlömilch-Rouché, 270
  in integral form, 355
  in Peano's form, 274
  of a series, 427
  in Taylor's formula, 270
Remarkable limit
  the first, 22, 127
  the second, 23, 128
Repeated limiting value, 484
Riemann's theorem on rearrangement of series, 446
Rolle's theorem, 256
Root of a polynomial, 201
  multiple, single, 204
Rouché-Schlömilch form of remainder, 270
Rule of comparison of real numbers, 41

Secant, 26
Sequence
  bounded, 58, 481
  convergent, 63, 479
  fundamental, 82, 480
  infinitely large, 59
  infinitesimal, 59
  iterative, 403
  limit of, 63, 479
  limit inferior of, 78
  limit superior of, 78
  limit point of, 75
  monotonic, 68
  number, 57
  unbounded, 59
Series
  absolutely convergent, 443
  alternating, 453
  comparison tests for, 430
  conditionally convergent, 441
  convergent, 424
  divergent, 424
  generalized harmonic, 433
  harmonic, 428
  Leibniz, 453
  number, 424
  partial sums of, 424
  sums of, 424
Set
  bounded, 43, 84
  boundary point of, 477
  closed, 477
  connected, 478
  convex, 536
  countable, 86
  dense in itself, 53
  diameter of, 491
  equivalent, 85
  finite, infinite, 85
  interior point of, 477
  isomorphic, 621
  limit point of (upper, lower), 621
  of the power of continuum, 85
  of real numbers, 43
Simpson rule (method of parabolas), 420
Singularity
  of a curve, 595
  of a surface, 568
Slope of a line, 27
Space
  coordinate, 473, 475
  Euclidean, 473, 475
  metric, 475
Squareable plane figure, 380
Stolz' theorem, 86
Strongly convex function, 542
Strophoid, 366
Subinterval, maximum length of, 325
Subsequence, 76
Sum
  lower, 328

of real numbers, 47
of a series, 424
upper, 328
Summation methods, 468

Tabular method of representing a function, 19, 96
Tangency of curves, 596
Tangent method (Newton's method), 400, 405
Tangent plane, 495
Tangent of the graph of a function, 26
Taylor's formula, 269, 517, 520
Trapezoids formula, 417
Trigonometric function, 121
inverse, 124

Undetermined coefficients, method of, 216
Undetermined multipliers, Lagrange's method of, 586
Upper sum, 328

Variable quantity, 17, 93
Vector function, 154
Velocity
average, 19
instantaneous, 19
Vertical asymptote, 313
Volume of a body (upper, lower), 386

Wallis's formula, 463
Weierstrass'
first theorem, 250, 490
second theorem, 263, 491
Weierstrass-Bolzano theorem, 80, 481

Zero point of a polynomial, 201

A BOOK OF PROBLEMS IN ORDINARY DIFFERENTIAL EQUATIONS

M. Krasnov, Cand.Sc., A. Kiselev, Cand.Sc.
and **G. Makarenko, Cand.Sc.**

This problem book contains exercises for courses in differential equations at technical institutes. Particular attention is given to the method of isoclines for equations of the first and second order, to problems of finding orthogonal trajectories, to the use of the method of superposition in solving linear differential equations of order n, to linear dependence and linear independence of a system of functions, to problems of solving linear equations with constant and variable coefficients, to boundary value problems for differential equations, to integrating equations in power series, to asymptotic integration, to integrating systems of differential equations, to Lyapunov stability, and to the operator method.

Each section begins with a summary of basic facts and comprises worked-out examples of typical problems.

The book contains 967 problems a considerable portion of which have been composed by the authors themselves. There are some exercises of a theoretical nature.

The book is intended for students of technical institutes.

A BRIEF COURSE OF HIGHER MATHEMATICS

V. Kudryavtsev, D.Sc. and
B. Demidovich, D.Sc.

This textbook was written as a course in higher mathematics for university students majoring in various fields of natural sciences.
The text consists of twenty six chapters and two appendices. Each chapter contains a large number of test questions and exercises to make the subject matter more readily comprehensible. Answers are provided for all problems in the main text, and every effort has been made to ensure that they are accurate.
Translated from the fifth (1978) Russian edition.

VECTOR ANALYSIS

M. Krasnov, Cand.Sc., A. Kiselev, Cand.Sc.
and **G. Makarenko, Cand.Sc.**

This text is designed for students of engineering colleges and also for practicing engineers who feel a need to refresh their knowledge of such an important area of higher mathematics as vector analysis. Each section of the text starts out with a brief review of the essentials of theory (propositions, definitions, formulas). This is followed by detailed solutions of examples and problems.
The book contains about 300 problems to be solved by the student. They are all provided with hints and answers. The 41 drawings that accompany the text serve to help the reader in analyzing the theoretical material and problems.

Professor ILYIN Vladimir Aleksandrovich is Head of a Chair at Moscow State University, senior research worker at the Institute of Mathematics of the USSR Academy of Sciences, Lomonosov Prize and USSR State Prize winner. He is the author of a number of textbooks for universities and more than a hundred scientific papers on the theory of partial differential equations, on the theory of functions and functional analysis and on the spectral theory of differential operators.

His most important scientific achievements concern the establishment of exact conditions for the expansion of elliptic operators in eigenfunctions and of exact conditions for convergence of spectral expansions corresponding to non-self-adjoint differential operators, and also the search for conditions for solvability of boundary-value and mixed problems for differential operators with nonsmooth coefficients and nonsmooth boundary manifolds.

Professor POZNYAK Eduard Genrikhovich, D.Sc. (Phys. and Math.) is a lecturer of mathematics at the Physics Department of Moscow State University, USSR State Prize Winner. He is the author of a number of University text-books and more than 70 scientific papers dealing with various divisions of geometry and differential equations related to geometry.

His most important results are connected with the theory of isometric em-bedding in the whole of two-dimensional Riemannian manifolds of negative curvature in Euclidean spaces, investigations connected with Lobachevsky's geometry and with various divisions of natural science.